

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
25 October 2001 (25.10.2001)

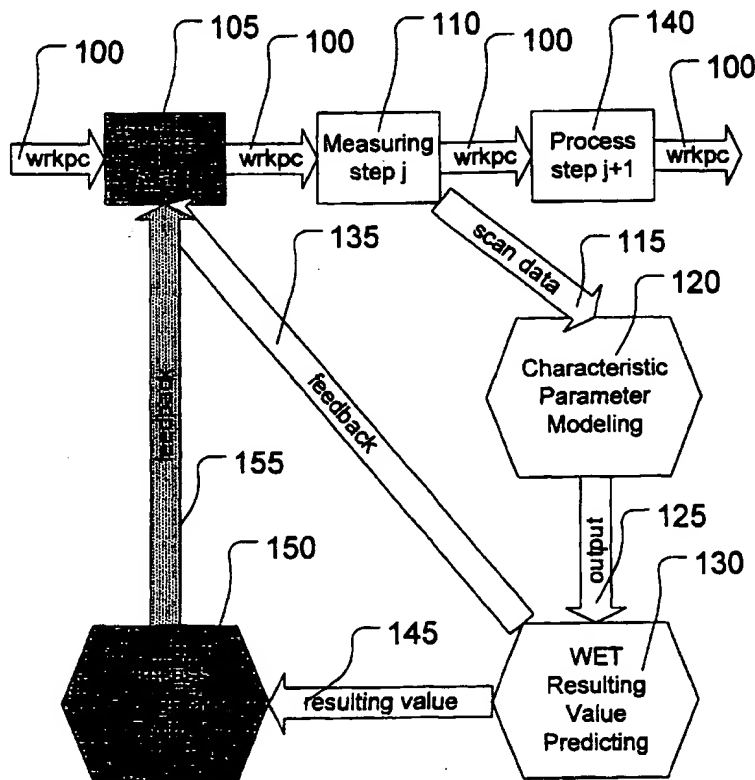
PCT

(10) International Publication Number
WO 01/80306 A2

- (51) International Patent Classification⁷: **H01L 21/66** (72) Inventors: **TOPRAC, Anthony, John**; 4023 Walnut Clay, Austin, TX 78731 (US). **MILLER, Michael, L.**; 2614 Little Elm Trail, Cedar Park, TX 78613 (US).
- (21) International Application Number: **PCT/US01/01562**
- (22) International Filing Date: 16 January 2001 (16.01.2001) (74) Agent: **DRAKE, Paul, S.**; Advanced Micro Devices, Inc., M/S 562, 5204 East Ben White Boulevard, Austin, TX 78741 (US).
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data: 09/548,779 13 April 2000 (13.04.2000) US (81) Designated States (*national*): JP, KR.
- (71) Applicant: **ADVANCED MICRO DEVICES, INC.** [US/US]; One AMD Place, Mail Stop 68, Sunnyvale, CA 94088-3453 (US). Published: (84) Designated States (*regional*): European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).
- without international search report and to be republished upon receipt of that report

[Continued on next page]

(54) Title: AUTOMATED PROCESS MONITORING AND ANALYSIS SYSTEM FOR SEMICONDUCTOR PROCESSING



(57) Abstract: A method is provided for manufacturing, the method comprising processing a workpiece (100), measuring a parameter (110) characteristic of the processing, and forming an output signal (125) corresponding to the characteristic parameter (110) measured by using the characteristic parameter (110) measured as an input to a transistor model (120). The method also comprises predicting (130) a wafer electrical test (WET) resulting value (145) based on the output signal (125), detecting (150) faulty processing based on the predicted WET resulting value (145), and correcting (135, 155) the faulty processing.

WO 01/80306 A2



For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

AUTOMATED PROCESS MONITORING AND ANALYSIS SYSTEM FOR SEMICONDUCTOR PROCESSING

TECHNICAL FIELD

5 This invention relates generally to semiconductor fabrication technology, and, more particularly, to a method for semiconductor fabrication process monitoring and analysis.

BACKGROUND ART

 There is a constant drive within the semiconductor industry to increase the quality, reliability and throughput of integrated circuit devices, *e.g.*, microprocessors, memory devices, and the like. This drive is fueled
10 by consumer demands for higher quality computers and electronic devices that operate more reliably. These demands have resulted in a continual improvement in the manufacture of semiconductor devices, *e.g.*, transistors, as well as in the manufacture of integrated circuit devices incorporating such transistors. Additionally, reducing defects in the manufacture of the components of a typical transistor also lowers the overall cost per transistor as well as the cost of integrated circuit devices incorporating such transistors.

15 The technologies underlying semiconductor processing tools have attracted increased attention over the last several years, resulting in substantial refinements. However, despite the advances made in this area, many of the processing tools that are currently commercially available suffer certain deficiencies. In particular, such tools often lack advanced process data monitoring capabilities, such as the ability to provide historical parametric data in a user-friendly format, as well as event logging, real-time graphical display of both current processing parameters
20 and the processing parameters of the entire run, and remote, *i.e.*, local site and worldwide, monitoring. These deficiencies can engender nonoptimal control of critical processing parameters, such as throughput accuracy, stability and repeatability, processing temperatures, mechanical tool parameters, and the like. This variability manifests itself as within-run disparities, run-to-run disparities and tool-to-tool disparities that can propagate into deviations in product quality and performance, whereas an ideal monitoring and diagnostics system for such tools
25 would provide a means of monitoring this variability, as well as providing means for optimizing control of critical parameters.

 Among the parameters it would be useful to monitor and control are critical dimensions (CDs) and doping levels for transistors (and other semiconductor devices), as well as overlay errors in photolithography. CDs are the smallest feature sizes that particular processing devices may be capable of producing. For example, the minimum
30 widths w of polycrystalline (polysilicon or poly) gate lines for metal oxide semiconductor field effect transistors (MOSFETs or MOS transistors) may correspond to one critical dimension (CD) for a semiconductor device having such transistors. Similarly, the junction depth d_j (depth below the surface of a doped substrate to the bottom of a heavily doped source/drain region formed within the doped substrate) may be another critical dimension (CD) for a semiconductor device such as an MOS transistor. Doping levels may depend on dosages of ions implanted into the
35 semiconductor devices, the dosages typically being given in numbers of ions per square centimeter at ion implant energies typically given in keV.

 However, traditional statistical process control (SPC) techniques are often inadequate to control precisely CDs and doping levels in semiconductor and microelectronic device manufacturing so as to optimize device performance and yield. Typically, SPC techniques set a target value, and a spread about the target value, for the
40 CDs, doping levels, and/or overlay errors in photolithography. The SPC techniques then attempt to minimize the deviation from the target value without automatically adjusting and adapting the respective target values to

optimize the semiconductor device performance, as measured by wafer electrical test (WET) measurement characteristics, for example, and/or to optimize the semiconductor device yield and throughput. Furthermore, blindly minimizing non-adaptive processing spreads about target values may not increase processing yield and throughput.

5 Traditional control techniques are frequently ineffective in reducing off-target processing and in improving sort yields. For example, the wafer electrical test (WET) measurements are typically not performed on processed wafers until quite a long time after the wafers have been processed, sometimes not until weeks later. When one or more of the processing steps are producing resulting wafers that WET measurements indicate are unacceptable, causing the resulting wafers to be scrapped, this misprocessing goes undetected and uncorrected for
10 quite a while, often for weeks, leading to many scrapped wafers, much wasted material and decreased overall throughput. Similarly, process and/or tool problems throughout the wafer processing are typically not analyzed fast enough, and final wafer yields are not evaluated on a die-by-die basis. Furthermore, data sets for making correlations between processing and/or tool trace data, on the one hand, and testing data, such as WET measurements, on the other, are typically manually extracted by the process engineers and put together, a very
15 time-consuming procedure.

 The present invention is directed to overcoming, or at least reducing the effects of, one or more of the problems set forth above.

DISCLOSURE OF INVENTION

 In one aspect of the present invention, a method is provided for manufacturing, the method comprising
20 processing a workpiece, measuring a parameter characteristic of the processing, and forming an output signal corresponding to the characteristic parameter measured by using the characteristic parameter measured as an input to a transistor model. The method also comprises predicting a wafer electrical test (WET) resulting value based on the output signal, detecting faulty processing based on the predicted WET resulting value, and correcting the faulty processing.

25 In another aspect of the present invention, a computer-readable, program storage device is provided, encoded with instructions that, when executed by a computer, perform a method for manufacturing a workpiece, the method comprising processing the workpiece, measuring a parameter characteristic of the processing, and forming an output signal corresponding to the characteristic parameter measured by using the characteristic parameter measured as an input to a transistor model. The method also comprises predicting a wafer electrical test
30 (WET) resulting value based on the output signal, detecting faulty processing based on the predicted WET resulting value, and correcting the faulty processing.

 In yet another aspect of the present invention, a computer programmed to perform a method of manufacturing is provided, the method comprising processing a workpiece, measuring a parameter characteristic of the processing, and forming an output signal corresponding to the characteristic parameter measured by using the
35 characteristic parameter measured as an input to a transistor model. The method also comprises predicting a wafer electrical test (WET) resulting value based on the output signal, detecting faulty processing based on the predicted WET resulting value, and correcting the faulty processing.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention may be understood by reference to the following description taken in conjunction with the accompanying drawings, in which the leftmost significant digit(s) in the reference numerals denote(s) the first figure in which the respective reference numerals appear, and in which:

5 Figures 1-14 schematically illustrate various embodiments of a method for manufacturing according to the present invention; and, more particularly:

 Figures 1-2 and 5-9 schematically illustrate a flow chart for various embodiments of a method for manufacturing according to the present invention;

10 Figures 3-4 schematically illustrate critical dimension (CD) measurements of features formed on a workpiece and an MOS transistor representative of MOS transistors tested in various embodiments of a method for manufacturing according to the present invention;

 Figure 10 schematically illustrates a method for fabricating a semiconductor device practiced in accordance with the present invention;

15 Figure 11 schematically illustrates workpieces being processed using a MOSFET processing tool, using a plurality of control input signals, in accordance with the present invention;

 Figures 12-13 schematically illustrate one particular embodiment of the process and tool in Figure 11; and

 Figure 14 schematically illustrates one particular embodiment of the method of Figure 10 as may be practiced with the process and tool of Figures 12-13.

20 While the invention is susceptible to various modifications and alternative forms, specific embodiments thereof have been shown by way of example in the drawings and are herein described in detail. It should be understood, however, that the description herein of specific embodiments is not intended to limit the invention to the particular forms disclosed, but on the contrary, the intention is to cover all modifications, equivalents, and alternatives falling within the spirit and scope of the invention as defined by the appended claims.

MODE(S) FOR CARRYING OUT THE INVENTION

25 Illustrative embodiments of the invention are described below. In the interest of clarity, not all features of an actual implementation are described in this specification. It will of course be appreciated that in the development of any such actual embodiment, numerous implementation-specific decisions must be made to achieve the developers' specific goals, such as compliance with system-related and business-related constraints, which will vary from one implementation to another. Moreover, it will be appreciated that such a development effort might be complex and time-consuming, but would nevertheless be a routine undertaking for those of
30 ordinary skill in the art having the benefit of this disclosure.

 Illustrative embodiments of a method for manufacturing according to the present invention are shown in Figures 1-14. As shown in Figure 1, a workpiece 100, such as a semiconducting substrate or wafer, having one or more process layers and/or semiconductor devices such as an MOS transistor disposed thereon, for example, is
35 delivered to a processing step j 105, where j may have any value from $j = 1$ to $j = N$. The total number N of processing steps, such as masking, etching, depositing material and the like, used to form the finished workpiece 100, may range from $N = 1$ to about any finite value.

 As shown in Figure 2, the workpiece 100 is sent from the processing step j 105 and delivered to a measuring step j 110. In the measuring step j 110, the workpiece 100 is measured by having a metrology or
40 measuring tool (not shown) measure one or more parameters characteristic of the processing performed in any of the previous processing steps (such as processing step j 105, where j may have any value from $j = 1$ to $j = N$). The

measurements in the measuring step j 110 produce scan data 115 indicative of the one or more characteristic parameters measured in the measuring step j 110. As shown in Figure 2, if there is further processing to do on the workpiece 100 (if $j < N$), then the workpiece 100 may be sent from the measuring step j 110 and delivered to a processing step j+1 140 for further processing, and then sent on from the processing step j+1 140.

5 In various illustrative embodiments, there is further processing to do on the workpiece 100 ($j < N$) and the measuring step j 110 may involve a critical dimension (CD) measurement of a structure formed on the workpiece 100. Figure 3 schematically illustrates the critical dimension (CD) measurement of a gate structure 300 formed on the workpiece 100. As shown in Figure 3, a gate dielectric 310 for the gate structure 300 (for an MOS transistor 400 as shown in Figure 4) may be formed above a structure layer 305, such as a semiconducting substrate 10 (e.g., a silicon wafer). The gate dielectric 310 may be formed by a variety of known techniques for forming such layers, e.g., chemical vapor deposition (CVD), low-pressure CVD (LPCVD), plasma-enhanced CVD (PECVD), thermal growth (such as substrate oxidation in a furnace), and the like, and may have a thickness ranging from approximately 20-200 Å, for example. The gate dielectric 310 may be formed from a variety of dielectric materials and may, for example, be an oxide (e.g., Ge oxide), a nitride (e.g., GaAs nitride), an oxynitride (e.g., GaP 15 oxynitride), silicon dioxide (SiO_2), a nitrogen-bearing oxide (e.g., nitrogen-bearing SiO_2), a nitrogen-doped oxide (e.g., N_2 -implanted SiO_2), silicon nitride (Si_3N_4), silicon oxynitride ($\text{Si}_x\text{O}_y\text{N}_z$), and the like. In one illustrative embodiment, the gate dielectric 310 is comprised of a silicon dioxide (SiO_2) having a thickness of approximately 50 Å, which is formed by an LPCVD process for higher throughput.

As shown in Figure 3, a polycrystalline silicon or poly gate conductive layer 310 for the gate structure 300 20 (for the MOS transistor 400 as shown in Figure 4) may be formed above the gate dielectric 310. The poly gate conductive layer 310 may be formed by a variety of known techniques for forming such layers, e.g., CVD, LPCVD, PECVD, sputtering, physical vapor deposition (PVD), and the like, and may have a thickness ranging from approximately 500-5000 Å. In one illustrative embodiment, the poly gate conductive layer 310 has a thickness of approximately 2000 Å and is formed by an LPCVD process for higher throughput. The poly gate 25 conductive layer 310 and the gate dielectric 310 together may constitute the gate structure 300.

As shown in Figure 3, the measuring step j 110 may involve the critical dimension (CD) measurement of the width W of the gate structure 300. The width W of the gate structure 300 may be related to the channel length L of the MOS transistor 400 as shown in Figure 4. Alternatively, as shown in Figure 4, the measuring step j 110 may involve the critical dimension (CD) measurement of a poly gate conductive layer 310 thickness t_p of the MOS 30 transistor 400. In various other alternative embodiments, the measuring step j 110 may involve other measurements such as a spacer 425 width w_s , a silicide (such as TiSi_2) 435 thickness t_s , and/or a gate dielectric 310 thickness t_{ox} , for example. The parameter and/or parameters measured in the measuring step j 110 may be characteristic of the processing performed on the workpiece 105 in the processing step j 105.

As shown in Figure 4, a metal oxide semiconductor field effect transistor (MOSFET or MOS 35 transistor) 400 may be formed on the semiconducting substrate 305, such as doped-silicon. The MOS transistor 400 may have the doped-poly gate 310 formed above the gate dielectric 315 formed above the semiconducting substrate 305. The doped-poly gate 310 and the gate dielectric 315 may be separated from N^+ -doped (P^+ -doped) source/drain regions 420 of the MOS transistor 400 by dielectric spacers 425. The dielectric spacers 425 may be formed above N^+ -doped (P^+ -doped) source/drain extension (SDE) regions 430.

40 The N^+ -doped (P^+ -doped) source/drain extension (SDE) regions 430 are typically provided to reduce the magnitude of the maximum channel electric field found close to the N^+ -doped (P^+ -doped) source/drain regions 420

of the MOS transistor 400, and, thereby, to reduce the associated hot-carrier effects. The lower (or lighter) doping of the N⁻-doped (P⁻-doped) source/drain extension (SDE) regions 430, relative to the higher (or heavier) doping of the N⁺-doped (P⁺-doped) source/drain regions 420 of the MOS transistor 400, reduces the magnitude of the maximum channel electric field found close to the N⁺-doped (P⁺-doped) source/drain regions 420 of the MOS transistor 400, but increases the source-to-drain resistances of the N⁻-doped (P⁻-doped) source/drain extension (SDE) regions 430.

A titanium (Ti) metal layer (not shown) may have been blanket-deposited on the MOS transistor 400 and then subjected to an initial rapid thermal anneal (RTA) process performed at a temperature ranging from approximately 450-800°C for a time ranging from approximately 15-60 seconds. At surfaces 440 of active areas 445, such as the N⁺-doped (P⁺-doped) source/drain regions 420 and the doped-poly gate 310, exposed Si reacts upon heating with the Ti metal to form a titanium silicide (TiSi₂) layer 435 the surfaces 440 of the active areas 445. The Ti metal is not believed to react with the dielectric spacers 425 upon heating. A wet chemical strip of the Ti metal removes excess, unreacted portions (not shown) of the Ti metal layer 235, leaving behind the self-aligned silicided (salicided) TiSi₂ layer 435 only at and below the surfaces 440 of the active areas 445. The salicided TiSi₂ 435 may then be subjected to a final RTA process performed at a temperature ranging from approximately 800-1100°C for a time ranging from approximately 10-60 seconds.

As shown in Figure 4, the MOS transistor 400 may be specified by several processing parameters. For example, the doped-poly gate 310 may have a width W that, in turn, determines a channel length L . The channel length L is the distance between the two metallurgical N⁻-P (P⁻-N) junctions formed below the gate dielectric 315 for an N-MOS (P-MOS) transistor 400, the two metallurgical N⁻-P (P⁻-N) junctions being between the N⁻-doped (P⁻-doped) source/drain extension (SDE) regions 430 and the semiconducting substrate 305. Further, another junction (having a junction depth d_j) below the N⁺-doped (P⁺-doped) source/drain regions 420 may be formed between the N⁺-doped (P⁺-doped) source/drain regions 420 and the semiconducting substrate 305. The semiconducting substrate 305 may have a doping level N_D (N_A) reflecting the density of donor (acceptor) impurities typically being given in numbers of ions per square centimeter for an N-type (P-type) semiconducting substrate 305. In addition, the N⁺-doped (P⁺-doped) source/drain regions 420 and the N⁻-doped (P⁻-doped) source/drain extension (SDE) regions 430 may each have respective doping levels N_{D+} and N_{D-} (N_{A+} and N_{A-}). The respective doping levels may depend on dosages of ions implanted into the N⁺-doped (P⁺-doped) source/drain regions 420 and the N⁻-doped (P⁻-doped) source/drain extension (SDE) regions 430, the dosages typically being given in numbers of ions per square centimeter at ion implant energies typically given in keV. Further, the gate dielectric 315 may have a thickness t_{ox} .

As shown in Figure 5, the scan data 115 is sent from the measuring step j 110 and delivered to a characteristic parameter modeling step 120. In the characteristic parameter modeling step 120, the one or more characteristic parameters measured in the measuring step j 110 may be input into a characteristic parameter model. The characteristic parameter model may map the one or more characteristic parameters measured in the measuring step j 110 onto one or more parameters that specify the completed workpiece 100. For example, the characteristic parameter model may be a transistor model. Delivering the scan data 115 to the characteristic parameter model in the characteristic parameter modeling step 120 produces an output signal 125.

As shown in Figure 6, the output signal 125 is sent from the characteristic parameter modeling step 120 and delivered to a wafer electrical test (WET) resulting value predicting step 130, producing at least one WET resulting value 145. In the WET resulting value predicting step 130, the characteristic parameter model may be

used to predict one or more of the WET resulting value(s) 145 that would result if the semiconductor device and/or devices and/or process layers formed on the workpiece 100 were subjected to WET measurements in eventual WET steps performed later, sometimes weeks later. The WET may measure current and/or voltage responses of MOS transistors formed on the workpiece 100, for example, and/or capacitances and/or resistances of elements of MOS transistors formed on the workpiece 100.

For example, a WET measurement of a cobalt silicided (CoSi_2) polysilicon serpentine structure (not shown) may be predicted, before the WET measurement is actually performed, by a characteristic parameter model with inputs from the relevant processing steps. The inputs from the relevant processing steps may comprise, but are not limited to, the critical dimension (CD) measurements of the width and thickness of the polysilicon of the cobalt silicided (CoSi_2) polysilicon serpentine structure, the thickness of the cobalt (Co) deposited thereon, and parametrics associated with the rapid thermal annealing process used to form the cobalt silicide (CoSi_2), such as the input power, measured temperature, and gas flows. Another example may be a WET measurement of transistor structure. In this case, the WET measurement may be a measurement of the drive current through a test transistor (like the MOS transistor 400, as shown in Figure 4). This drive current measurement may be predicted by a characteristic parameter model, before the WET measurement is actually performed, using inputs from data gathered during the relevant processing steps. In this case, the inputs from the relevant processing steps may comprise, but are not limited to, implant dose and energies, critical dimension (CD) measurement of a poly gate conductive layer 310 thickness t_p , spacer 425 width w_s , silicide (such as TiSi_2) 435 thickness t_s , and/or a gate dielectric 310 thickness t_{ox} , for example. The width W of the gate structure 300 may be related to the channel length L of the MOS transistor 400 as shown in Figure 4.

In various illustrative embodiments, characteristic parameters y_α , $\alpha = 1$ to $\alpha = m$, obtained using in-line process metrology, may be mapped to predicted WET resulting values x_β , $\beta = 1$ to $\beta = n$, in the completed workpiece 100 by the mapping $T^{-1}(y_\alpha) = x_\beta$. The characteristic parameters y_α , $\alpha = 1$ to $\alpha = m$, may be represented as m vectors each having s components, or, equivalently, as an $s \times m$ matrix $Y_{s \times m}$, whose m columns are the m

vectors y_α , $\alpha = 1$ to $\alpha = m$: $Y_{s \times m} = (y_\alpha) = (y_1 \cdots y_m) = (y_{\beta\alpha}) = \begin{pmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{s1} & \cdots & y_{sm} \end{pmatrix}$. Similarly, the

predicted WET resulting values x_β , $\beta = 1$ to $\beta = n$, may be represented as n vectors each having t components, or, equivalently, as an $t \times n$ matrix $X_{t \times n}$, whose n columns are the n vectors x_β , $\beta = 1$ to $\beta = n$:

$X_{t \times n} = (x_\alpha) = (x_1 \cdots x_n) = (x_{\beta\alpha}) = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{t1} & \cdots & x_{tn} \end{pmatrix}$. In various illustrative embodiments, the mapping

$T^{-1}(y_\alpha) = x_\beta$, may be represented as multiplication of the $s \times m$ matrix $Y_{s \times m}$ by the $t \times s$ matrix $L_{t \times s}$ on the left and by the $m \times n$ matrix $R_{m \times n}$ on the right:

$$L_{t \times s} Y_{s \times m} R_{m \times n} = X_{t \times n} = \begin{pmatrix} l_{11} & \cdots & l_{1s} \\ \vdots & \ddots & \vdots \\ l_{t1} & \cdots & l_{ts} \end{pmatrix} \begin{pmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{s1} & \cdots & y_{sm} \end{pmatrix} \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{t1} & \cdots & x_{tn} \end{pmatrix}$$

In various illustrative embodiments, the mapping $T^{-1}(y_\alpha) = x_\beta$ of the characteristic parameters y_α , $\alpha = 1$ to $\alpha = m$, obtained using in-line process metrology, onto the predicted WET resulting values x_β , $\beta = 1$ to $\beta = n$, in the

completed workpiece 100, may be determined by using partial least squares (PLS) techniques. The partial least squares (PLS) techniques attempt to decompose both the characteristic parameters y_{α} , $\alpha = 1$ to $\alpha = m$, obtained using in-line process metrology, and the predicted WET resulting values \hat{x}_{β} , $\beta = 1$ to $\beta = n$, in the completed workpiece 100, each into a set of "scores" and "loadings." The scores may represent the relationship between samples (for example, drifts in the values from one sample to another). The loadings may show the relationships between variables (for example, the relationship of one WET parameter to another). In the partial least squares (PLS) techniques, the relationship of the loadings U_k for the characteristic parameters y_{α} , $\alpha = 1$ to $\alpha = m$, and the loadings P_k for the predicted WET resulting values \hat{x}_{β} , $\beta = 1$ to $\beta = n$, is linear: $U_k = T^{-1}P_k$. Using historical measurements of the characteristic parameters y_{α} , $\alpha = 1$ to $\alpha = m$, and the predicted WET resulting values \hat{x}_{β} , $\beta = 1$ to $\beta = n$, an optimal set of scores, loadings, and the mapping T^{-1} may be determined.

The mapping $T^{-1}(y_{\alpha}) = \hat{x}_{\beta}$ of the characteristic parameters y_{α} , $\alpha = 1$ to $\alpha = m$, obtained using in-line process metrology, onto the predicted WET resulting values \hat{x}_{β} , $\beta = 1$ to $\beta = n$, in the completed workpiece 100, may be used on-line to detect and/or correct errant processing that might cause the completed workpiece 100 to be consigned to WET scrap, thereby reducing wasted material and increasing throughput of corrected completed workpieces 100. For example, in various illustrative embodiments, the mapping $T^{-1}(y_{\alpha}) = \hat{x}_{\beta}$ may be inverted $y_{\alpha} = T(\hat{x}_{\beta})$ to define one or more changes in the processing performed in any of the previous processing steps (such as processing step j 105, where j may have any value from $j = 1$ to $j = N$) that need to be made to bring the one or more characteristic parameter values y_{α} , $\alpha = 1$ to $\alpha = m$, measured in the measuring step j 110 within a range of specification values.

The prediction of the WET resulting value(s) 145 (based on the output signal 125) in the WET resulting value predicting step 130 may be used to alert an engineer of the need to adjust the processing performed in any of the previous processing steps (such as processing step j 105, where j may have any value from $j = 1$ to $j = N$). The engineer may also alter, for example, the type of characteristic parameter modeled in the characteristic parameter modeling step 120, affecting the output signal 125 produced.

As shown in Figure 7, a feedback control signal 135 may be sent from the WET resulting value predicting step 130 to the processing step j 105 to adjust automatically the processing performed in the processing step j 105. In various alternative illustrative embodiments (not shown), the feedback control signal 135 may be sent from the WET resulting value predicting step 130 to any of the previous processing steps (similar to processing step j 105, where j may have any value from $j = 1$ to $j = N$) to adjust automatically the processing performed in any of the previous processing steps.

As shown in Figure 8, in addition to, and/or instead of, the feedback control signal 135, the WET resulting value(s) 145 may be sent from the WET resulting value predicting step 130 to a process change and control step 150. In the process change and control step 150, the WET resulting value(s) 145 may be used in a high-level supervisory control loop and/or used to detect faulty processing performed in any of the previous processing steps (such as processing step j 105, where j may have any value from $j = 1$ to $j = N$). Thereafter, as shown in Figure 10, a feedback control signal 155 may be sent from the process change and control step 150 to the processing step j 105 to adjust and/or correct the processing performed in the processing step j 105. In various alternative illustrative embodiments (not shown), the feedback control signal 155 may be sent from the process change and control step 150 to any of the previous processing steps (similar to processing step j 105, where j may have any value from $j = 1$ to $j = N$) to adjust and/or correct the processing performed in any of the previous processing steps.

The WET measurements of the semiconductor device and/or devices and/or process layers formed on the workpiece 100 that are performed in an eventual WET measuring step may measure current and/or voltage responses of the MOS transistors 400 formed on the workpiece 100, for example, and/or capacitances and/or resistances of elements of the MOS transistors 400 formed on the workpiece 100. Examples of WET transistor measurement(s) may include, but are not limited to, measurement(s) of threshold voltage(s) and/or source/drain drive current(s). Resistance measurement(s) at WET may include determination of intrinsic material sheet resistance and/or measurement(s) through a serpentine test structure and/or series resistance measurements on contact structures. Capacitance measurement(s) at WET may include measurements of the capacitance of the gate dielectric.

For example, the WET of the MOS transistors 400 formed on the workpiece 100 may measure the drain-source current I_D at different values of the drain voltage V_D , gate voltage V_G and/or substrate voltage (or bias) V_{BS} . By measuring change in the drain-source current I_D with change in the drain voltage V_D , at constant gate

voltage V_G , the channel conductance g_D may be determined from $g_D = \left. \frac{\partial I_D}{\partial V_D} \right|_{V_G = \text{const.}} = \frac{Z}{L} \mu_n C_i (V_G - V_T)$,

where Z is the channel width (in the direction perpendicular to the plane of the MOS transistor 400 in Figure 4), μ_n is the mobility of the electrons (related to the drift velocity v_{ndrift} of the electrons by $v_{ndrift} = \mu_n E$, where $E = V_D/L$ is the electric field across the drain/source), C_i is the capacitance per unit area ($C_i = \epsilon_{ox}/t_{ox}$, where $\epsilon_{ox} \approx 4$ is the dielectric constant for the gate dielectric 315), and V_T is the threshold voltage of the MOS transistor 400. Similarly, by measuring change in the drain-source current I_D with change in the gate voltage V_G , at constant drain voltage

V_D , the transconductance g_m may be determined from $g_m = \left. \frac{\partial I_D}{\partial V_G} \right|_{V_D = \text{const.}} = \frac{Z}{L} \mu_n C_i V_D$. Here, the linear region

of drain-source current I_D versus drain voltage V_D is used, where $I_D \approx \left(\frac{Z}{L} \right) \mu_n C_i (V_G - V_T) V_D$, for

$V_D \ll (V_G - V_T)$, and the threshold voltage V_T is given by $V_T = 2\psi_B + \frac{\sqrt{2\epsilon_s q N_A (2\psi_B)}}{C_i}$, where ψ_B is the

potential difference between the Fermi level E_F in the doped-poly gate 310 and the intrinsic (flat-band) Fermi level E_{Fi} in the P-type semiconducting substrate 305, ϵ_s is the dielectric constant for the P-type semiconducting substrate 305, q is the absolute value of the electric charge on an electron ($q = 1.60218 \times 10^{-19}$ Coulombs), and the doping level N_A reflects the density of acceptor impurities for the P-type semiconducting substrate 305.

The WET measurements, represented generally by a vector \mathbf{x} (here $\beta = n = 1$ for \mathbf{x}_β), such as those given above, may be put into an MOS transistor model, represented generally by a function $T(\mathbf{x})$, which maps the WET measurements \mathbf{x} into a set of parameters, represented generally by a vector \mathbf{y} (here $\alpha = m = 1$ for \mathbf{y}_α), characteristic of the processing performed in at least one of the processing steps j 105, where j may have any value from $j = 1$ to $j = N$, so that $T(\mathbf{x}) = \mathbf{y}$. The transistor model may be inverted, represented generally by a function $T^{-1}(\mathbf{y}) = \mathbf{x}$, which maps the characteristic processing parameters \mathbf{y} into the WET measurements \mathbf{x} .

For example, one illustrative embodiment of an MOS transistor model function $T(\mathbf{x})$ gives the minimum channel length L_{min} (related to the doped-poly gate 310 width W) for which long-channel subthreshold behavior can be observed. In this illustrative embodiment, the MOS transistor model function $T(\mathbf{x})$ gives the minimum channel

length L_{\min} by the simple empirical relation: $L_{\min} = 0.4[d_j t_{ox} (W_s + W_D)^2]^{1/3}$, measured in μm , where the junction depth d_j is measured in μm , the gate dielectric 315 thickness t_{ox} is the numerical value of the number of Å units (so the dimensions work out), and $(W_s + W_D)$ is the sum of the source and drain depletion depths, respectively, also measured in μm . In a one-dimensional abrupt junction formulation, the source depletion depth W_s may be

5 given by: $W_s = \sqrt{\frac{2\epsilon_s}{qN_A}(V_{bi} + V_{BS})}$ and the drain depletion depth W_D may be given by:

$$W_D = \sqrt{\frac{2\epsilon_s}{qN_A}(V_D + V_{bi} + V_{BS})}, \text{ where } V_{bi} \text{ is the built-in voltage of the junction.}$$

Another illustrative embodiment of an MOS transistor model function $T(x)$ gives the minimum channel length L_{\min} by the more complicated empirical relation:

$$L_{\min} = Af_1(\delta V_T / \delta V_D) [f_2(t_{ox}) + B[f_3(W_s + W_D) + C[f_4(d_j) + D]]], \text{ where the functions } f_i, \text{ for } i = 1, 2, 3, 4,$$

10 and the constants A, B, C, D , may be determined by fitting this equation for the minimum channel length L_{\min} to device simulations. For example, $f_1(\delta V_T / \delta V_D) = (\delta V_T / \delta V_D)^{-0.37}$, $f_2(t_{ox}) = t_{ox}$, $f_3(W_s + W_D) = W_s + W_D$, $f_4(d_j) = d_j$, $A = 2.2 \mu\text{m}^{-2}$, $B = 0.012 \mu\text{m}$, $C = 0.15 \mu\text{m}$, and $D = 2.9 \mu\text{m}$ appear to give a good fit. In this illustrative embodiment, the inverted MOS transistor model function $T^{-1}(y)$ gives the variation $(\delta V_T / \delta V_D)$ of the threshold voltage V_T with the drain voltage V_D , for example, by the more complicated empirical relation:

15 $\delta V_T / \delta V_D = f_1^{-1}(L_{\min} / \{A[f_2(t) + B[f_3(W_s + W_D) + C[f_4(d_j) + D]]\})$. For the fit where $f_1(\delta V_T / \delta V_D) = (\delta V_T / \delta V_D)^{-0.37}$, $f_1^{-1}(y) = (y)^{-1/(0.37)}$, for example.

In various illustrative embodiments, partial least squares (PLS) modeling may be used to effect the mapping $T^{-1}(\underline{y}_\alpha) = \underline{x}_\beta$ of the characteristic parameters \underline{y}_α , $\alpha = 1$ to $\alpha = m$, obtained using in-line process metrology, onto the predicted WET resulting values \underline{x}_β , $\beta = 1$ to $\beta = n$, in the completed workpiece 100. In various alternative

20 illustrative embodiments, Principal Components Analysis (PCA) modeling may be used to effect the mapping $T^{-1}(\underline{y}_\alpha) = \underline{x}_\beta$ of the characteristic parameters \underline{y}_α , $\alpha = 1$ to $\alpha = m$, obtained using in-line process metrology, onto the predicted WET resulting values \underline{x}_β , $\beta = 1$ to $\beta = n$, in the completed workpiece 100.

In various illustrative embodiments, the engineer may be provided with advanced process data monitoring capabilities, such as the ability to provide historical parametric data in a user-friendly format, as well as event

25 logging, real-time graphical display of both current processing parameters and the processing parameters of the entire run, and remote, i.e., local site and worldwide, monitoring. These capabilities may engender more optimal control of critical processing parameters, such as throughput accuracy, stability and repeatability, processing temperatures, mechanical tool parameters, and the like. This more optimal control of critical processing parameters reduces this variability. This reduction in variability manifests itself as fewer within-run disparities, fewer

30 run-to-run disparities and fewer tool-to-tool disparities. This reduction in the number of these disparities that can propagate means fewer deviations in product quality and performance. In such an illustrative embodiment of a method of manufacturing according to the present invention, a monitoring and diagnostics system may be provided that monitors this variability and optimizes control of critical parameters.

Figure 10 illustrates one particular embodiment of a method 1000 practiced in accordance with the present

35 invention. Figure 11 illustrates one particular apparatus 1100 with which the method 1000 may be practiced. For the sake of clarity, and to further an understanding of the invention, the method 1000 shall be disclosed in the

context of the apparatus 1100. However, the invention is not so limited and admits wide variation, as is discussed further below.

Referring now to both Figures 10 and 11, a batch or lot of workpieces or wafers 1105 is being processed through a MOSFET processing tool 1110. The MOSFET processing tool 1110 may be any MOSFET processing tool known to the art, such as an ion implanter, a process layer deposition and/or etching tool, a photolithography tool, and the like, provided it includes the requisite control capabilities. The MOSFET processing tool 1110 includes a MOSFET processing tool controller 1115 for this purpose. The nature and function of the MOSFET processing tool controller 1115 will be implementation specific.

For instance, the MOSFET processing tool controller 1115 may control MOSFET processing control input parameters such as MOSFET processing recipe control input parameters. As shown in Figure 4, the MOS transistor 400 may be specified by several processing parameters. For example, the doped-poly gate 310 may have a width w that, in turn, determines a channel length L . The channel length L is the distance between the two metallurgical N-P (P-N) junctions formed below the gate dielectric 315 for an N-MOS (P-MOS) transistor 400, the two metallurgical N-P (P-N) junctions being between the N-doped (P-doped) source/drain extension (SDE) regions 430 and the semiconducting substrate 305. The doped-poly gate 310 may have a thickness t_p , the spacer 425 may have a width w_s , a silicide (such as cobalt silicide, CoSi_2 , or titanium silicide, TiSi_2) 435 may have a thickness t_s , and the gate dielectric 310 may have a thickness t_{ox} , for example. Further, another junction (having a junction depth d_j) below the N⁺-doped (P⁺-doped) source/drain regions 420 may be formed between the N⁺-doped (P⁺-doped) source/drain regions 420 and the semiconducting substrate 305. The semiconducting substrate 305 may have a doping level N_D (N_A) reflecting the density of donor (acceptor) impurities typically being given in numbers of ions per square centimeter for an N-type (P-type) semiconducting substrate 305. In addition, the N⁺-doped (P⁺-doped) source/drain regions 420 and the N-doped (P-doped) source/drain extension (SDE) regions 430 may each have respective doping levels N_{D+} and N_{D-} (N_{A+} and N_{A-}). The respective doping levels may depend on dosages of ions implanted into the N⁺-doped (P⁺-doped) source/drain regions 420 and the N-doped (P-doped) source/drain extension (SDE) regions 430, the dosages typically being given in numbers of ions per square centimeter at ion implant energies typically given in keV. Four workpieces 1105 are shown in Figure 11, but the lot of workpieces or wafers, i.e., the "wafer lot," may be any practicable number of wafers from one to any finite number.

The method 1000 begins, as set forth in box 1020, by measuring a parameter characteristic of the MOSFET processing performed on the workpiece 1105 in the MOSFET processing tool 1110. The nature, identity, and measurement of characteristic parameters will be largely implementation specific and even tool specific. For instance, capabilities for monitoring process parameters vary, to some degree, from tool to tool. Greater sensing capabilities may permit wider latitude in the characteristic parameters that are identified and measured and the manner in which this is done. Conversely, lesser sensing capabilities may restrict this latitude. For example, a gate poly etch MOSFET processing tool reads the gate critical dimension of a workpiece 1105, and/or an average of the gate critical dimensions of the workpieces 1105 in a lot, using a metrology tool (not shown). The gate critical dimension of a workpiece 1105, and/or an average of the gate critical dimensions of the workpieces 1105 in a lot, is an illustrative example of a parameter characteristic of the MOSFET processing performed on the workpiece in the MOSFET processing tool 1110.

Turning to Figure 11, in this particular embodiment, the MOSFET processing process characteristic parameters are measured and/or monitored by tool sensors (not shown). The outputs of these tool sensors are

transmitted to a computer system 1130 over a line 1120. The computer system 1130 analyzes these sensor outputs to identify the characteristic parameters.

Returning, to Figure 10, once the characteristic parameter is identified and measured, the method 1000 proceeds by modeling the measured and identified characteristic parameter, using a wafer electrical test (WET) prediction model, as set forth in box 1030. The computer system 1130 in Figure 11 is, in this particular embodiment, programmed to model the characteristic parameter. The manner in which this modeling occurs will be implementation specific.

In the embodiment of Figure 11, a database 1135 stores a plurality of wafer electrical test (WET) prediction models that might potentially be applied, depending upon which characteristic parameter is measured. This particular embodiment, therefore, requires some *a priori* knowledge of the characteristic parameters that might be measured. The computer system 1130 then extracts an appropriate wafer electrical test (WET) prediction model from the database 1135 of potential models to apply to the measured characteristic parameters. If the database 1135 does not include an appropriate wafer electrical test (WET) prediction model, then the characteristic parameter may be ignored, or the computer system 1130 may attempt to develop one, if so programmed. The database 1135 may be stored on any kind of computer-readable, program storage medium, such as an optical disk 1140, a floppy disk 1145, or a hard disk drive (not shown) of the computer system 1130. The database 1135 may also be stored on a separate computer system (not shown) that interfaces with the computer system 1130.

Modeling of the measured characteristic parameter may be implemented differently in alternative embodiments. For instance, the computer system 1130 may be programmed using some form of artificial intelligence to analyze the sensor outputs and controller inputs to develop a wafer electrical test (WET) prediction model on-the-fly in a real-time implementation. This approach might be a useful adjunct to the embodiment illustrated in Figure 11, and discussed above, where characteristic parameters are measured and identified for which the database 1135 has no appropriate wafer electrical test (WET) prediction model.

The method 1000 of Figure 10 then proceeds by applying the wafer electrical test (WET) prediction model to modify a MOSFET processing control input parameter, as set forth in box 1040. Depending on the implementation, applying the wafer electrical test (WET) prediction model may yield either a new value for the MOSFET processing control input parameter or a correction to the existing MOSFET processing control input parameter. The new MOSFET processing control input is then formulated from the value yielded by the wafer electrical test (WET) prediction model and is transmitted to the MOSFET processing tool controller 1115 over the line 1120. The MOSFET processing tool controller 1115 then controls subsequent MOSFET processing process operations in accordance with the new MOSFET processing control inputs.

Some alternative embodiments may employ a form of feedback to improve the modeling of characteristic parameters. The implementation of this feedback is dependent on several disparate facts, including the tool's sensing capabilities and economics. One technique for doing this would be to monitor at least one effect of the model's implementation and update the model based on the effect(s) monitored. The update may also depend on the model. For instance, a linear model may require a different update than would a non-linear model, all other factors being the same.

As is evident from the discussion above, some features of the present invention are implemented in software. For instance, the acts set forth in the boxes 1020-1040 in Figure 10 are, in the illustrated embodiment, software-implemented, in whole or in part. Thus, some features of the present invention are implemented as instructions encoded on a computer-readable, program storage medium. The program storage medium may be of

any type suitable to the particular implementation. However, the program storage medium will typically be magnetic, such as the floppy disk 1145 or the computer 1130 hard disk drive (not shown), or optical, such as the optical disk 1140. When these instructions are executed by a computer, they perform the disclosed functions. The computer may be a desktop computer, such as the computer 1130. However, the computer might alternatively be a processor embedded in the MOSFET processing tool 1110. The computer might also be a laptop, a workstation, or a mainframe in various other embodiments. The scope of the invention is not limited by the type or nature of the program storage medium or computer with which embodiments of the invention might be implemented.

Thus, some portions of the detailed descriptions herein are, or may be, presented in terms of algorithms, functions, techniques, and/or processes. These terms enable those skilled in the art most effectively to convey the substance of their work to others skilled in the art. These terms are here, and are generally, conceived to be a self-consistent sequence of steps leading to a desired result. The steps are those requiring physical manipulations of physical quantities. Usually, though not necessarily, these quantities take the form of electromagnetic signals capable of being stored, transferred, combined, compared, and otherwise manipulated.

It has proven convenient at times, principally for reasons of common usage, to refer to these signals as bits, values, elements, symbols, characters, terms, numbers, and the like. All of these and similar terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities and actions. Unless specifically stated otherwise, or as may be apparent from the discussion, terms such as "processing," "computing," "calculating," "determining," "displaying," and the like, used herein refer to the action(s) and processes of a computer system, or similar electronic and/or mechanical computing device, that manipulates and transforms data, represented as physical (electromagnetic) quantities within the computer system's registers and/or memories, into other data similarly represented as physical quantities within the computer system's memories and/or registers and/or other such information storage, transmission and/or display devices.

Construction of an Illustrative Apparatus. An exemplary embodiment 1200 of the apparatus 1100 in Figure 11 is illustrated in Figures 12-13, in which the apparatus 1200 comprises a portion of an Advanced Process Control (APC) system. Figures 12-13 are conceptualized, structural and functional block diagrams, respectively, of the apparatus 1200. A set of processing steps is performed on a lot of wafers 1205 on a MOSFET processing tool 1210. Because the apparatus 1200 is part of an advanced process control (APC) system, the wafers 1205 are processed on a run-to-run basis. Thus, process adjustments are made and held constant for the duration of a run, based on run-level measurements or averages. A "run" may be a lot, a batch of lots, or even an individual wafer.

In this particular embodiment, the wafers 1205 are processed by the MOSFET processing tool 1210 and various operations in the process are controlled by a plurality of MOSFET processing control input signals on a line 1220 between the MOSFET processing tool 1210 and a workstation 1230. Exemplary MOSFET processing control inputs for this embodiment might include those for the gate critical dimension (width and/or thickness), the source/drain junction depth, doping profiles, spacer width, silicide thickness, gate dielectric thickness, and the like.

As described above, and as shown in Figure 4, the MOS transistor 400 may be specified by several processing parameters. For example, the doped-poly gate 310 may have a width w that, in turn, determines a channel length L . The channel length L is the distance between the two metallurgical N-P (P-N) junctions formed below the gate dielectric 315 for an N-MOS (P-MOS) transistor 400, the two metallurgical N-P (P-N) junctions being between the N-doped (P-doped) source/drain extension (SDE) regions 430 and the semiconducting substrate 305. The doped-poly gate 310 may have a thickness t_p , the spacer 425 may have a width w_s , a silicide (such as cobalt silicide, CoSi_2 , or titanium silicide, TiSi_2) 435 may have a thickness t_s , and the gate dielectric 310

may have a thickness t_{ox} for example. Further, another junction (having a junction depth d_j) below the N⁺-doped (P⁺-doped) source/drain regions 420 may be formed between the N⁺-doped (P⁺-doped) source/drain regions 420 and the semiconducting substrate 305. The semiconducting substrate 305 may have a doping level N_D (N_A) reflecting the density of donor (acceptor) impurities typically being given in numbers of ions per square centimeter for an N-type (P-type) semiconducting substrate 305. In addition, the N⁺-doped (P⁺-doped) source/drain regions 420 and the N-doped (P-doped) source/drain extension (SDE) regions 430 may each have respective doping levels N_{D+} and N_{D-} (N_{A+} and N_{A-}). The respective doping levels may depend on dosages of ions implanted into the N⁺-doped (P⁺-doped) source/drain regions 420 and the N-doped (P-doped) source/drain extension (SDE) regions 430, the dosages typically being given in numbers of ions per square centimeter at ion implant energies typically given in keV.

When a process step in the MOSFET processing tool 1210 is concluded, the semiconductor wafers 1205 being processed in the MOSFET processing tool 1210 are examined in a review station 1217. The MOSFET processing control inputs generally affect the characteristic parameters of the semiconductor wafers 1205 and, hence, the variability and properties of the acts performed by the MOSFET processing tool 1210 on the wafers 1205. Once errors are determined from the examination after the run of a lot of wafers 1205, the MOSFET processing control inputs on the line 1220 are modified for a subsequent run of a lot of wafers 1205. Modifying the control signals on the line 1220 is designed to improve the next process step in the MOSFET processing tool 1210. The modification is performed in accordance with one particular embodiment of the method 1000 set forth in Figure 10, as described more fully below. Once the relevant MOSFET processing control input signals for the MOSFET processing tool 1210 are updated, the MOSFET processing control input signals with new settings are used for a subsequent run of semiconductor devices.

Referring now to both Figures 12 and 13, the MOSFET processing tool 1210 communicates with a manufacturing framework comprising a network of processing modules. One such module is an advanced process control (APC) system manager 1340 resident on the computer 1240. This network of processing modules constitutes the advanced process control (APC) system. The MOSFET processing tool 1210 generally includes an equipment interface 1310 and a sensor interface 1315. A machine interface 1330 resides on the workstation 1230. The machine interface 1330 bridges the gap between the advanced process control (APC) framework, *e.g.*, the advanced process control (APC) system manager 1340, and the equipment interface 1310. Thus, the machine interface 1330 interfaces the MOSFET processing tool 1210 with the advanced process control (APC) framework and supports machine setup, activation, monitoring, and data collection. The sensor interface 1315 provides the appropriate interface environment to communicate with external sensors such as LabView® or other sensor bus-based data acquisition software. Both the machine interface 1330 and the sensor interface 1315 use a set of functionalities (such as a communication standard) to collect data to be used. The equipment interface 1310 and the sensor interface 1315 communicate over the line 1220 with the machine interface 1330 resident on the workstation 1230.

More particularly, the machine interface 1330 receives commands, status events, and collected data from the equipment interface 1310 and forwards these as needed to other advanced process control (APC) components and event channels. In turn, responses from advanced process control (APC) components are received by the machine interface 1330 and rerouted to the equipment interface 1310. The machine interface 1330 also reformats and restructures messages and data as necessary. The machine interface 1330 supports the startup/shutdown procedures within the advanced process control (APC) System Manager 1340. It also serves as an advanced

process control (APC) data collector, buffering data collected by the equipment interface 1310, and emitting appropriate data collection signals.

In the particular embodiment illustrated, the advanced process control (APC) system is a factory-wide software system, but this is not necessary to the practice of the invention. The control strategies taught by the present invention can be applied to virtually any semiconductor MOSFET processing tool on a factory floor. Indeed, the present invention may be simultaneously employed on multiple MOSFET processing tools in the same factory or in the same fabrication process. The advanced process control (APC) framework permits remote access and monitoring of the process performance. Furthermore, by utilizing the advanced process control (APC) framework, data storage can be more convenient, more flexible, and less expensive than data storage on local drives. However, the present invention may be employed, in some alternative embodiments, on local drives.

The illustrated embodiment deploys the present invention onto the advanced process control (APC) framework utilizing a number of software components. In addition to components within the advanced process control (APC) framework, a computer script is written for each of the semiconductor MOSFET processing tools involved in the control system. When a semiconductor MOSFET processing tool in the control system is started in the semiconductor manufacturing fab, the semiconductor MOSFET processing tool generally calls upon a script to initiate the action that is required by the MOSFET processing tool controller. The control methods are generally defined and performed using these scripts. The development of these scripts can comprise a significant portion of the development of a control system.

In this particular embodiment, there are several separate software scripts that perform the tasks involved in controlling the MOSFET processing operation. There is one script for the MOSFET processing tool 1210, including the review station 1217 and the MOSFET processing tool controller 1215. There is also a script to handle the actual data capture from the review station 1217 and another script that contains common procedures that can be referenced by any of the other scripts. There is also a script for the advanced process control (APC) system manager 1340. The precise number of scripts, however, is implementation specific and alternative embodiments may use other numbers of scripts.

Operation of an Illustrative Apparatus. Figure 14 illustrates one particular embodiment 1400 of the method 1000 in Figure 10. The method 1400 may be practiced with the apparatus 1200 illustrated in Figures 12-13, but the invention is not so limited. The method 1400 may be practiced with any apparatus that may perform the functions set forth in Figure 14. Furthermore, the method 1000 in Figure 10 may be practiced in embodiments alternative to the method 1400 in Figure 14.

Referring now to all of Figures 12-14, the method 1400 begins with processing a lot of wafers 1205 through MOSFET processing tools, such as the MOSFET processing tool 1210, as set forth in box 1410. In this particular embodiment, the MOSFET processing tool 1210 has been initialized for processing by the advanced process control (APC) system manager 1340 through the machine interface 1330 and the equipment interface 1310. In this particular embodiment, before the MOSFET processing tool 1210 is run, the advanced process control (APC) system manager script is called to initialize the MOSFET processing tool 1210. At this step, the script records the identification number of the MOSFET processing tool 1210 and the lot number of the wafers 1205. The identification number is then stored against the lot number in a data store 1260. The rest of the script, such as the APCData call and the Setup and StartMachine calls, are formulated with blank or dummy data in order to force the machine to use default settings.

As part of this initialization, the initial setpoints for MOSFET processing control are provided to the MOSFET processing tool controller 1215 over the line 1220. These initial setpoints may be determined and implemented in any suitable manner known to the art. In the particular embodiment illustrated, MOSFET processing controls are implemented by control threads. Each control thread acts like a separate controller and is differentiated by various process conditions. For MOSFET processing control, the control threads are separated by a combination of different conditions. These conditions may include, for example, the semiconductor MOSFET processing tool 1210 currently processing the wafer lot, the semiconductor product, the semiconductor manufacturing operation, and one or more of the semiconductor processing tools (not shown) that previously processed the semiconductor wafer lot.

Control threads are separated because different process conditions affect the MOSFET processing error differently. By isolating each of the process conditions into its own corresponding control thread, the MOSFET processing error can become a more accurate portrayal of the conditions in which a subsequent semiconductor wafer lot in the control thread will be processed. Since the error measurement is more relevant, changes to the MOSFET processing control input signals based upon the error will be more appropriate.

The control thread for the MOSFET processing control scheme depends upon the current MOSFET processing tool, current operation, the product code for the current lot, and the identification number at a previous processing step. The first three parameters are generally found in the context information that is passed to the script from the MOSFET processing tool 1210. The fourth parameter is generally stored when the lot is previously processed. Once all four parameters are defined, they are combined to form the control thread name; MOSP02_OPER01_PROD01_MOSP01 is an example of a control thread name. The control thread name is also stored in correspondence to the wafer lot number in the data store 1260.

Once the lot is associated with a control thread name, the initial settings for that control thread are generally retrieved from the data store 1260. There are at least two possibilities when the call is made for the information. One possibility is that there are no settings stored under the current control thread name. This can happen when the control thread is new, or if the information was lost or deleted. In these cases, the script initializes the control thread assuming that there is no error associated with it and uses the target values of the MOSFET processing errors as the MOSFET processing control input settings. It is preferred that the controllers use the default machine settings as the initial settings. By assuming some settings, the MOSFET processing errors can be related back to the control settings in order to facilitate feedback control.

Another possibility is that the initial settings are stored under the control thread name. In this case, one or more wafer lots have been processed under the same control thread name as the current wafer lot, and have also been measured for MOSFET processing error using the review station 1217. When this information exists, the MOSFET processing control input signal settings are retrieved from the data store 1260. These settings are then downloaded to the MOSFET processing tool 1210.

The wafers 1205 are processed through the MOSFET processing tool 1210. This may include, in the embodiment illustrated, any MOSFET processing known to the art, such as ion implantation, process layer deposition and/or etching, photolithography processing, and the like, provided it includes the requisite control capabilities. The wafers 1205 are measured on the review station 1217 after their MOSFET processing on the MOSFET processing tool 1210. The review station 1217 examines the wafers 1205 after they are processed for a number of errors. The data generated by the instruments of the review station 1217 is passed to the machine interface 1330 via sensor interface 1315 and the line 1220. The review station script begins with a number of

advanced process control (APC) commands for the collection of data. The review station script then locks itself in place and activates a data available script. This script facilitates the actual transfer of the data from the review station 1217 to the advanced process control (APC) framework. Once the transfer is completed, the script exits and unlocks the review station script. The interaction with the review station 1217 is then generally complete.

5 As will be appreciated by those skilled in the art having the benefit of this disclosure, the data generated by the review station 1217 should be preprocessed for use. Review stations, such as KLA review stations, provide the control algorithms for measuring the control error. Each of the error measurements, in this particular embodiment, corresponds to one of the MOSFET processing control input signals on the line 1220 in a direct manner. Before the error can be utilized to correct the MOSFET processing control input signal, a certain amount
10 of preprocessing is generally completed.

For example, preprocessing may include outlier rejection. Outlier rejection is a gross error check ensuring that the received data is reasonable in light of the historical performance of the process. This procedure involves comparing each of the MOSFET processing errors to its corresponding predetermined boundary parameter. In one embodiment, even if one of the predetermined boundaries is exceeded, the error data from the entire semiconductor
15 wafer lot is generally rejected.

To determine the limits of the outlier rejection, thousands of actual semiconductor manufacturing fabrication ("fab") data points are collected. The standard deviation for each error parameter in this collection of data is then calculated. In one embodiment, for outlier rejection, nine times the standard deviation (both positive and negative) is generally chosen as the predetermined boundary. This was done primarily to ensure that only the
20 points that are significantly outside the normal operating conditions of the process are rejected.

Preprocessing may also smooth the data, which is also known as filtering. Filtering is important because the error measurements are subject to a certain amount of randomness, such that the error significantly deviates in value. Filtering the review station data results in a more accurate assessment of the error in the MOSFET processing control input signal settings. In one embodiment, the MOSFET processing control scheme utilizes a
25 filtering procedure known as an Exponentially-Weighted Moving Average ("EWMA") filter, although other filtering procedures can be utilized in this context.

One embodiment for the EWMA filter is represented by Equation (1):

$$\text{AVG}_N = W * M_C + (1-W) * \text{AVG}_P \quad (1)$$

where

30 AVG_N = the new EWMA average;

W = a weight for the new average (AVG_N);

M_C = the current measurement; and

AVG_P = the previous EWMA average.

The weight is an adjustable parameter that can be used to control the amount of filtering and is generally
35 between zero and one. The weight represents the confidence in the accuracy of the current data point. If the measurement is considered accurate, the weight should be close to one. If there were a significant amount of fluctuations in the process, then a number closer to zero would be appropriate.

In one embodiment, there are at least two techniques for utilizing the EWMA filtering process. The first technique uses the previous average, the weight, and the current measurement as described above. Among the
40 advantages of utilizing the first implementation are ease of use and minimal data storage. One of the disadvantages of utilizing the first implementation is that this method generally does not retain much process information.

Furthermore, the previous average calculated in this manner would be made up of every data point that preceded it, which may be undesirable. The second technique retains only some of the data and calculates the average from the raw data each time.

5 The manufacturing environment in the semiconductor manufacturing fab presents some unique challenges. The order that the semiconductor wafer lots are processed through an MOSFET processing tool may not correspond to the order in which they are read on the review station. This could lead to the data points being added to the EWMA average out of sequence. Semiconductor wafer lots may be analyzed more than once to verify the error measurements. With no data retention, both readings would contribute to the EWMA average, which may be an undesirable characteristic. Furthermore, some of the control threads may have low volume, which may cause
10 the previous average to be outdated such that it may not be able to accurately represent the error in the MOSFET processing control input signal settings.

The MOSFET processing tool controller 1215, in this particular embodiment, uses limited storage of data to calculate the EWMA filtered error, *i.e.*, the first technique. Wafer lot data, including the lot number, the time the lot was processed, and the multiple error estimates, are stored in the data store 1260 under the control thread name.
15 When a new set of data is collected, the stack of data is retrieved from data store 1260 and analyzed. The lot number of the current lot being processed is compared to those in the stack. If the lot number matches any of the data present there, the error measurements are replaced. Otherwise, the data point is added to the current stack in chronological order, according to the time periods when the lots were processed. In one embodiment, any data point within the stack that is over 128 hours old is removed. Once the aforementioned steps are complete, the new filter average is calculated and stored to data store 1260.
20

Thus, the data is collected and preprocessed, and then processed to generate an estimate of the current errors in the MOSFET processing control input signal settings. First, the data is passed to a compiled Matlab® plug-in that performs the outlier rejection criteria described above. The inputs to a plug-in interface are the multiple error measurements and an array containing boundary values. The return from the plug-in interface is a single
25 toggle variable. A nonzero return denotes that it has failed the rejection criteria, otherwise the variable returns the default value of zero and the script continues to process.

After the outlier rejection is completed, the data is passed to the EWMA filtering procedure. The controller data for the control thread name associated with the lot is retrieved, and all of the relevant operation upon the stack of lot data is carried out. This includes replacing redundant data or removing older data. Once the
30 data stack is adequately prepared, it is parsed into ascending time-ordered arrays that correspond to the error values. These arrays are fed into the EWMA plug-in along with an array of the parameter required for its execution. In one embodiment, the return from the plug-in is comprised of the six filtered error values.

Returning to Figure 14, data preprocessing includes predicting the workpiece 1205 wafer electrical test (WET) measurement values that would be measured in a final wafer electrical test (WET) measurement step, using
35 a wafer electrical test (WET) model, as set forth in box 1420. Known, potential characteristic parameters may be identified by characteristic data patterns or may be identified as known consequences of modifications to MOSFET processing control. For example, the identification and modeling of how changes in gate critical dimension affect the predicted final wafer electrical test (WET) measurements may fall into this latter category.

The next step in the control process is to calculate the new settings for the MOSFET processing tool
40 controller 1215 of the MOSFET processing tool 1210. The previous settings for the control thread corresponding to the current wafer lot are retrieved from the data store 1260. This data is paired along with the current set of

MOSFET processing errors. The new settings are calculated by calling a compiled Matlab® plug-in. This application incorporates a number of inputs, performs calculations in a separate execution component, and returns a number of outputs to the main script. Generally, the inputs of the Matlab® plug-in are the MOSFET processing control input signal settings, the review station errors, an array of parameters that are necessary for the control algorithm, and a currently unused flag error. The outputs of the Matlab® plug-in are the new controller settings, calculated in the plug-in according to the controller algorithm described above.

A MOSFET processing process engineer or a control engineer, who generally determines the actual form and extent of the control action, can set the parameters. They include the threshold values, maximum step sizes, controller weights, and target values. Once the new parameter settings are calculated, the script stores the setting in the data store 1260 such that the MOSFET processing tool 1210 can retrieve them for the next wafer lot to be processed. The principles taught by the present invention can be implemented into other types of manufacturing frameworks.

Returning again to Figure 14, the calculation of new settings includes, as set forth in box 1430, modeling the workpiece 1205 WET values as a function of the MOSFET processing recipe parameters. This modeling may be performed by the Matlab® plug-in. In this particular embodiment, only known, potential characteristic parameters are modeled and the models are stored in a database 1235 accessed by a machine interface 1330. The database 1235 may reside on the workstation 1230, as shown, or some other part of the advanced process control (APC) framework. For instance, the models might be stored in the data store 1260 managed by the advanced process control (APC) system manager 1340 in alternative embodiments. The model will generally be a mathematical model, *i.e.*, an equation describing how the change(s) in MOSFET processing recipe control(s) affects the MOSFET processing performance and the WET measurements in the final WET, and the like. The transistor models, and/or processing step submodel(s), described in various illustrative embodiments given above are examples of such models.

The particular model used will be implementation specific, depending upon the particular MOSFET processing tool 1210 and the particular characteristic parameter being modeled. Whether the relationship in the model is linear or non-linear will be dependent on the particular parameters involved.

The new settings are then transmitted to and applied by the MOSFET processing tool controller 1215. Thus, returning now to Figure 14, once the workpiece 1205 WET values are modeled, the model is applied to modify at least one MOSFET processing recipe control input parameter, as set forth in box 1440. In this particular embodiment, the machine interface 1330 retrieves the model from the database 1235, plugs in the respective value(s), and determines the necessary change(s) in the MOSFET processing recipe control input parameter(s). The change is then communicated by the machine interface 1330 to the equipment interface 1310 over the line 1220. The equipment interface 1310 then implements the change.

The present embodiment furthermore provides that the models be updated. This includes, as set forth in boxes 1450-1460 of Figure 14, monitoring at least one effect of modifying the MOSFET processing recipe control input parameters (box 1450) and updating the applied model (box 1460) based on the effect(s) monitored. For instance, various aspects of the MOSFET processing tool 1210's operation will change as the MOSFET processing tool 1210 ages. By monitoring the effect of the MOSFET processing recipe change(s) implemented as a result of the characteristic parameter (*e.g.*, workpiece 1205 gate critical dimensions) measurement, the necessary value could be updated to yield superior performance.

As noted above, this particular embodiment implements an advanced process control (APC) system. Thus, changes are implemented "between" lots. The actions set forth in the boxes 1420-1460 are implemented after the current lot is processed and before the second lot is processed, as set forth in box 1470 of Figure 14. However, the invention is not so limited. Furthermore, as noted above, a lot may constitute any practicable number of wafers from one to several thousand (or practically any finite number). What constitutes a "lot" is implementation specific, and so the point of the fabrication process in which the updates occur will vary from implementation to implementation.

Any of the above-disclosed embodiments of a method of manufacturing according to the present invention enables the reduction of off-target processing and the improvement of sort yields. Additionally, any of the above-disclosed embodiments of a method of manufacturing according to the present invention enables semiconductor device fabrication with increased device accuracy and precision, increased efficiency and increased device yield, enabling a streamlined and simplified process flow, thereby decreasing the complexity and lowering the costs of the manufacturing process and increasing throughput.

The particular embodiments disclosed above are illustrative only, as the invention may be modified and practiced in different but equivalent manners apparent to those skilled in the art having the benefit of the teachings herein. Furthermore, no limitations are intended to the details of construction or design herein shown, other than as described in the claims below. It is therefore evident that the particular embodiments disclosed above may be altered or modified and all such variations are considered within the scope and spirit of the invention. Accordingly, the protection sought herein is as set forth in the claims below.

CLAIMS

1. A method of manufacturing, the method comprising:
processing a workpiece (100);
measuring a parameter (110) characteristic of the processing;
5 forming an output signal (125) corresponding to the characteristic parameter (110) measured by
using the characteristic parameter (110) measured as an input to a transistor
model (120);
predicting (130) a wafer electrical test (WET) resulting value (145) based on the output
signal (125);
10 detecting (150) faulty processing based on the predicted WET resulting value (145); and
correcting (135, 155) the faulty processing.
2. The method of claim 1, wherein forming the output signal (125) corresponding to the
characteristic parameter (110) measured by using the characteristic parameter (110) measured as the input to the
15 transistor model (120) comprises using the characteristic parameter (110) measured as the input to a partial least
squares transistor model (120), wherein using the characteristic parameter (110) measured as the input to the partial
least squares transistor model (120) comprises using the partial least squares transistor model (120) to map a set of
in-line process metrology input values to a set of WET measurement output values and wherein using the partial
least squares transistor model (120) to map the set of the in-line process metrology input values to the set of the
20 WET measurement output values comprises using the partial least squares transistor model (120) to define at least
a subset of the in-line process metrology input values having a significant effect on at least a subset of the WET
measurement output values.
3. The method of claim 1, wherein correcting (135, 155) the faulty processing comprises inverting
25 the transistor model (120) to define a change in the processing needed to bring subsequent predicted WET resulting
values (145) within a range of specification values.
4. The method of claim 2, wherein correcting (135, 155) the faulty processing comprises inverting
the transistor model (120) to define a change in the processing needed to bring subsequent predicted WET resulting
30 values (145) within a range of specification values.
5. The method of claim 3, wherein inverting the transistor model (120) to define the change in the
processing comprises defining a change in a critical dimension of a feature formed in the processing needed to
bring the subsequent predicted WET resulting values (145) within the range of specification values, wherein
35 defining the change in the critical dimension of the feature formed in the processing comprises defining the change
in the critical dimension of at least one of a poly gate line width, a spacer width, a gate dielectric thickness and a
silicide layer thickness of an MOS transistor.
6. The method of claim 3, wherein inverting the transistor model (120) to define the change in the
40 processing comprises defining a change in a doping level of a feature formed in the processing needed to bring the
subsequent predicted WET resulting values (145) within the range of specification values and defining the change

in the doping level of the feature formed in the processing comprises at least one of defining the change in the doping level of a source/drain region of an MOS transistor and defining the change in the doping level of a source/drain extension (SDE) region of an MOS transistor.

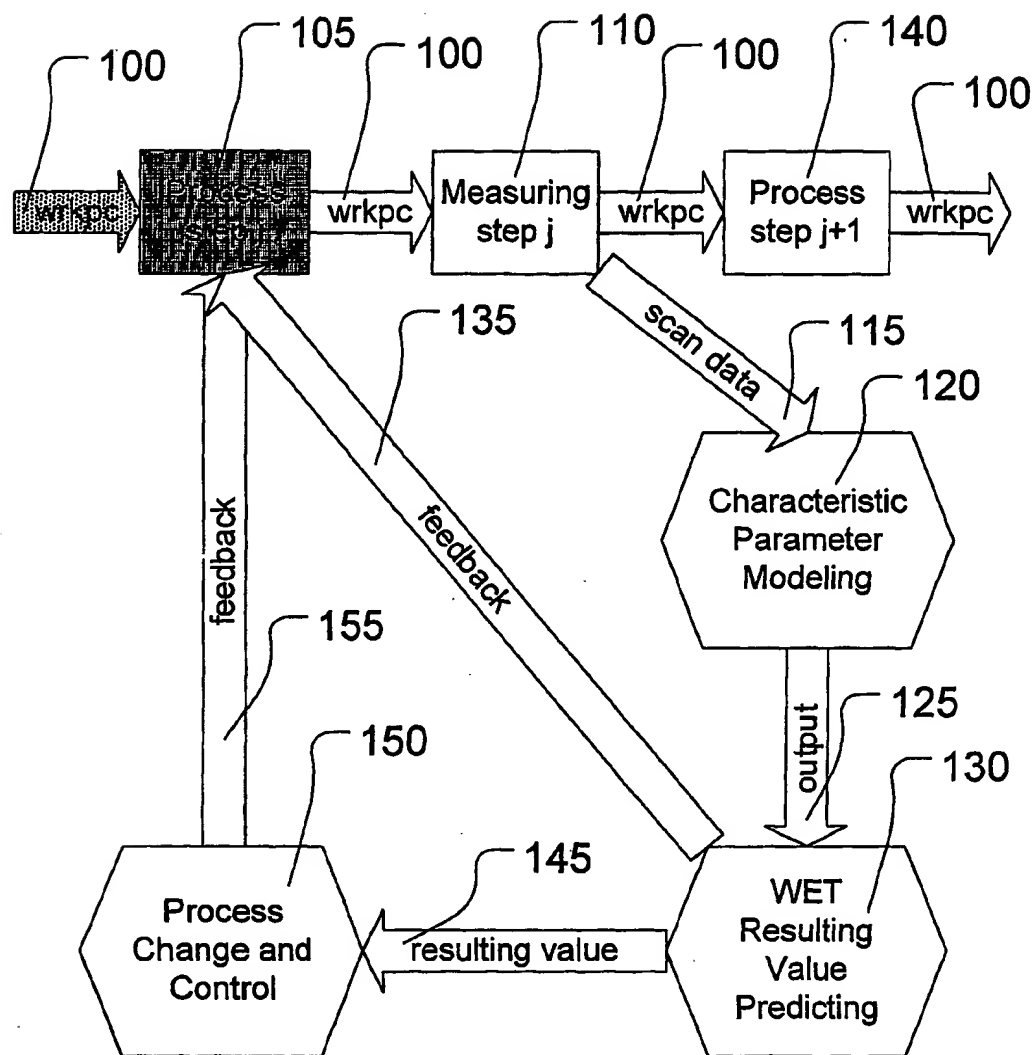
5 7. The method of claim 4, wherein inverting the transistor model (120) to define the change in the processing comprises defining a change in a critical dimension of a feature formed in the processing needed to bring the subsequent predicted WET resulting values (145) within the range of specification values, wherein defining the change in the critical dimension of the feature formed in the processing comprises defining the change in the critical dimension of at least one of a poly gate line width, a spacer width, a gate dielectric thickness and a
10 silicide layer thickness of an MOS transistor.

 8. The method of claim 4, wherein inverting the transistor model (120) to define the change in the processing comprises defining a change in a doping level of a feature formed in the processing needed to bring the subsequent predicted WET resulting values (145) within the range of specification values and defining the change
15 in the doping level of the feature formed in the processing comprises at least one of defining the change in the doping level of a source/drain region of an MOS transistor and defining the change in the doping level of a source/drain extension (SDE) region of an MOS transistor.

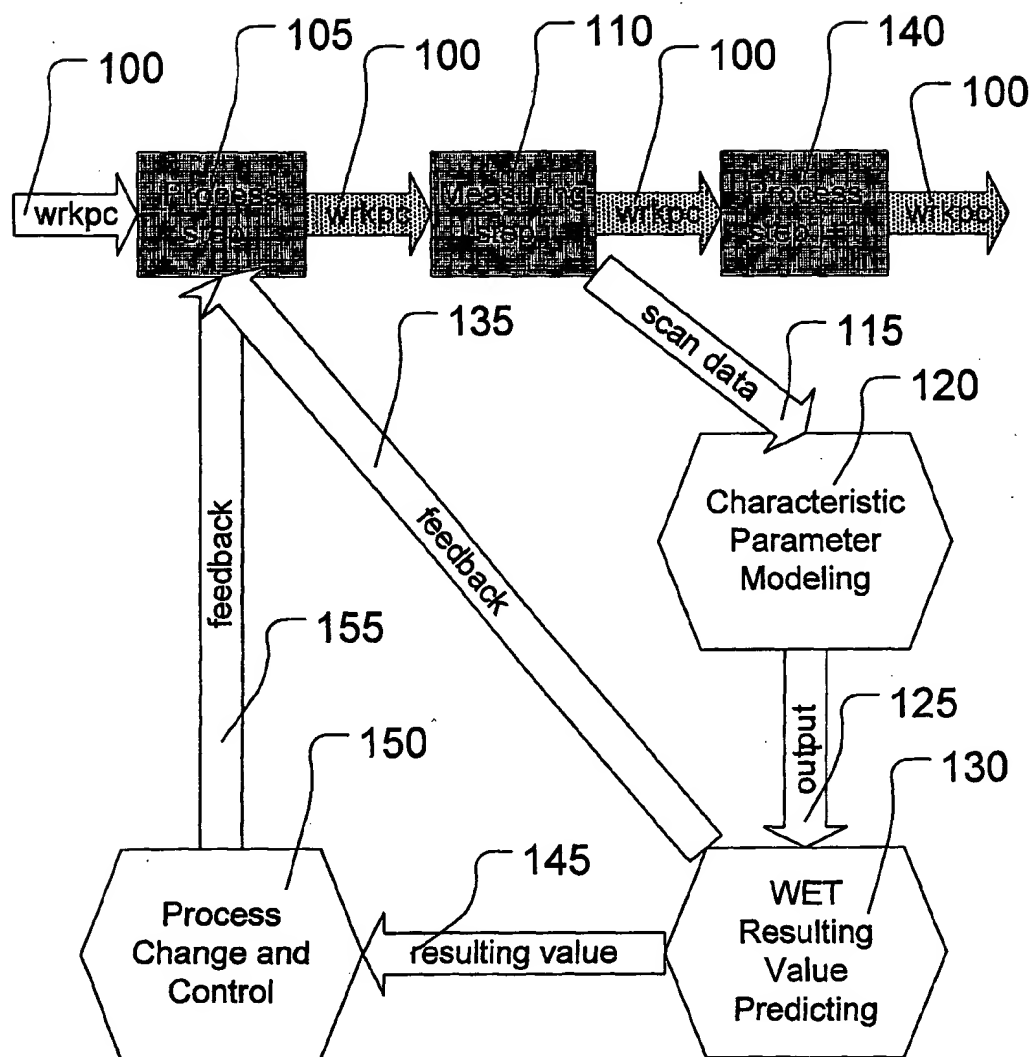
 9. A computer-readable, program storage device encoded with instructions that, when executed by a
20 computer, perform a method for manufacturing a workpiece (100), the method comprising:
 processing the workpiece (100);
 measuring a parameter (110) characteristic of the processing performed on the workpiece (100);
 forming an output signal (125) corresponding to the characteristic parameter (110) measured by
 using the characteristic parameter (110) measured as an input to a transistor
25 model (120);
 predicting (130) a wafer electrical test (WET) resulting value (145) based on the output
 signal (125);
 detecting (150) faulty processing based on the predicted WET resulting value (145); and
 correcting (135, 155) the faulty processing, wherein forming the output signal (125)
30 corresponding to the characteristic parameter (110) measured by using the characteristic parameter (110)
 measured as the input to the transistor model (120) comprises using the characteristic parameter (110)
 measured as the input to a partial least squares transistor model (120), wherein using the characteristic
 parameter (110) measured as the input to the partial least squares transistor model (120) comprises using
35 the partial least squares transistor model (120) to map a set of in-line process metrology input values to a
 set of WET measurement output values, wherein using the partial least squares transistor model (120) to
 map the set of the in-line process metrology input values to the set of the WET measurement output
 values comprises using the partial least squares transistor model (120) to define at least a subset of the
 in-line process metrology input values having a significant effect on at least a subset of the WET
 measurement output values and wherein correcting (135, 155) the faulty processing comprises inverting
40 the transistor model (120) to define a change in the processing needed to bring subsequent predicted WET
 resulting values (145) within a range of specification values.

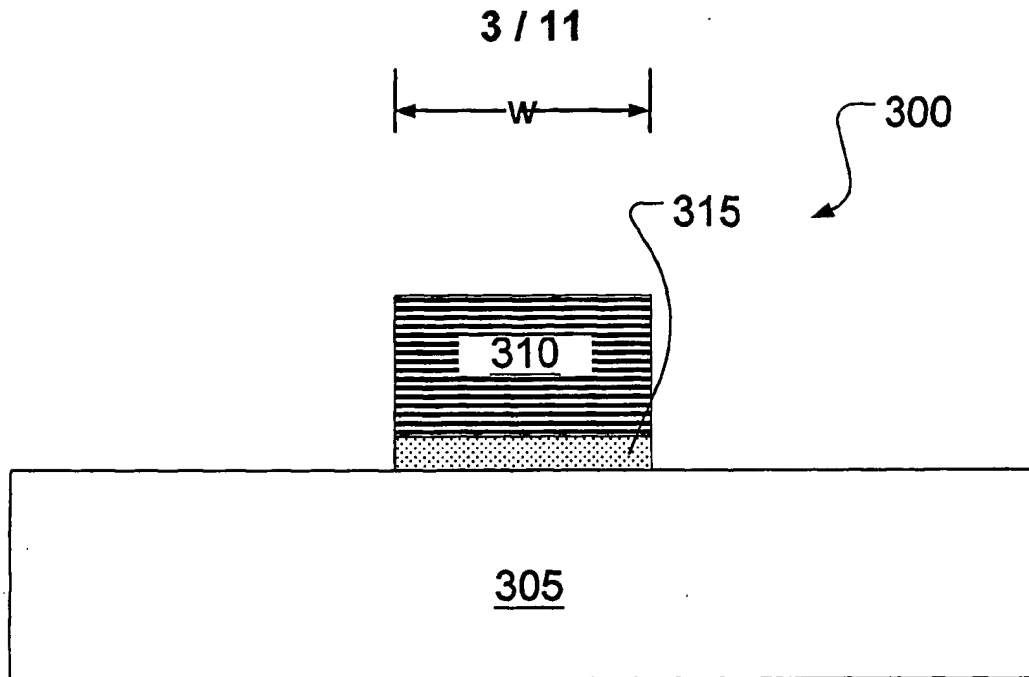
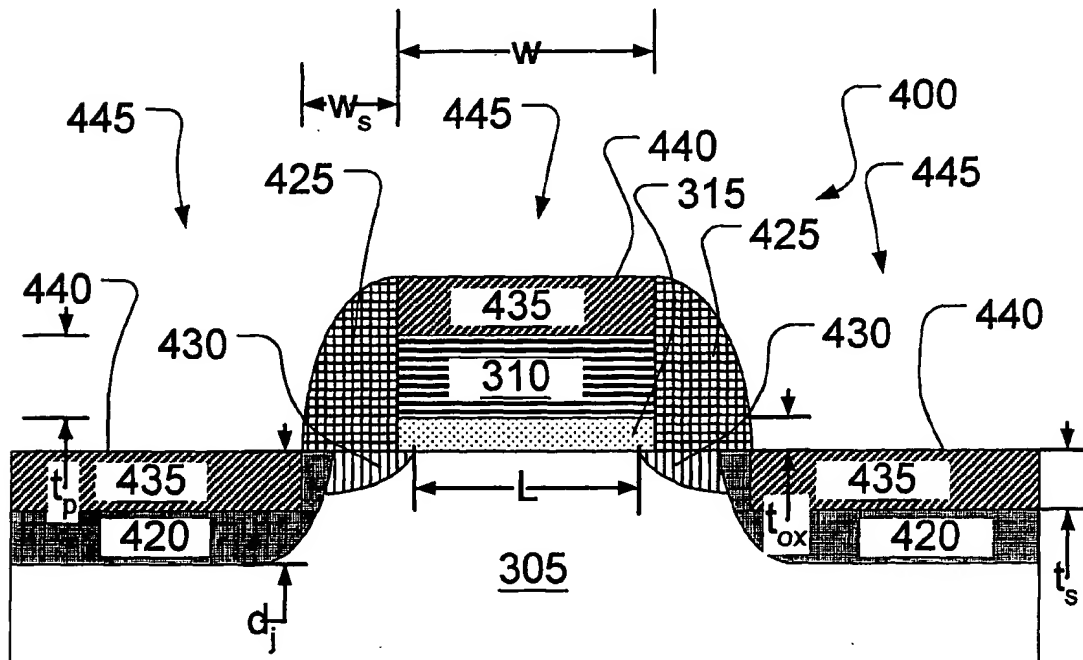
10. A computer programmed to perform a method of manufacturing, the method comprising:
- processing a workpiece (100);
- measuring a parameter (110) characteristic of the processing performed on the workpiece (100);
- forming an output signal (125) corresponding to the characteristic parameter (110) measured by
- 5 using the characteristic parameter (110) measured as an input to a transistor model (120);
- predicting (130) a wafer electrical test (WET) resulting value (145) based on the output signal (125);
- detecting (150) faulty processing based on the predicted WET resulting value (145); and
- 10 correcting (135, 155) the faulty processing, wherein forming the output signal (125) corresponding to the characteristic parameter (110) measured by using the characteristic parameter (110) measured as the input to the transistor model (120) comprises using the characteristic parameter (110) measured as the input to a partial least squares transistor model (120), wherein using the characteristic parameter (110) measured as the input to the partial least squares transistor model (120) comprises using
- 15 the partial least squares transistor model (120) to map a set of in-line process metrology input values to a set of WET measurement output values, wherein using the partial least squares transistor model (120) to map the set of the in-line process metrology input values to the set of the WET measurement output values comprises using the partial least squares transistor model (120) to define at least a subset of the in-line process metrology input values having a significant effect on at least a subset of the WET measurement output values and wherein correcting (135, 155) the faulty processing comprises inverting
- 20 the transistor model (120) to define a change in the processing needed to bring subsequent predicted WET resulting values (145) within a range of specification values.

1 / 11

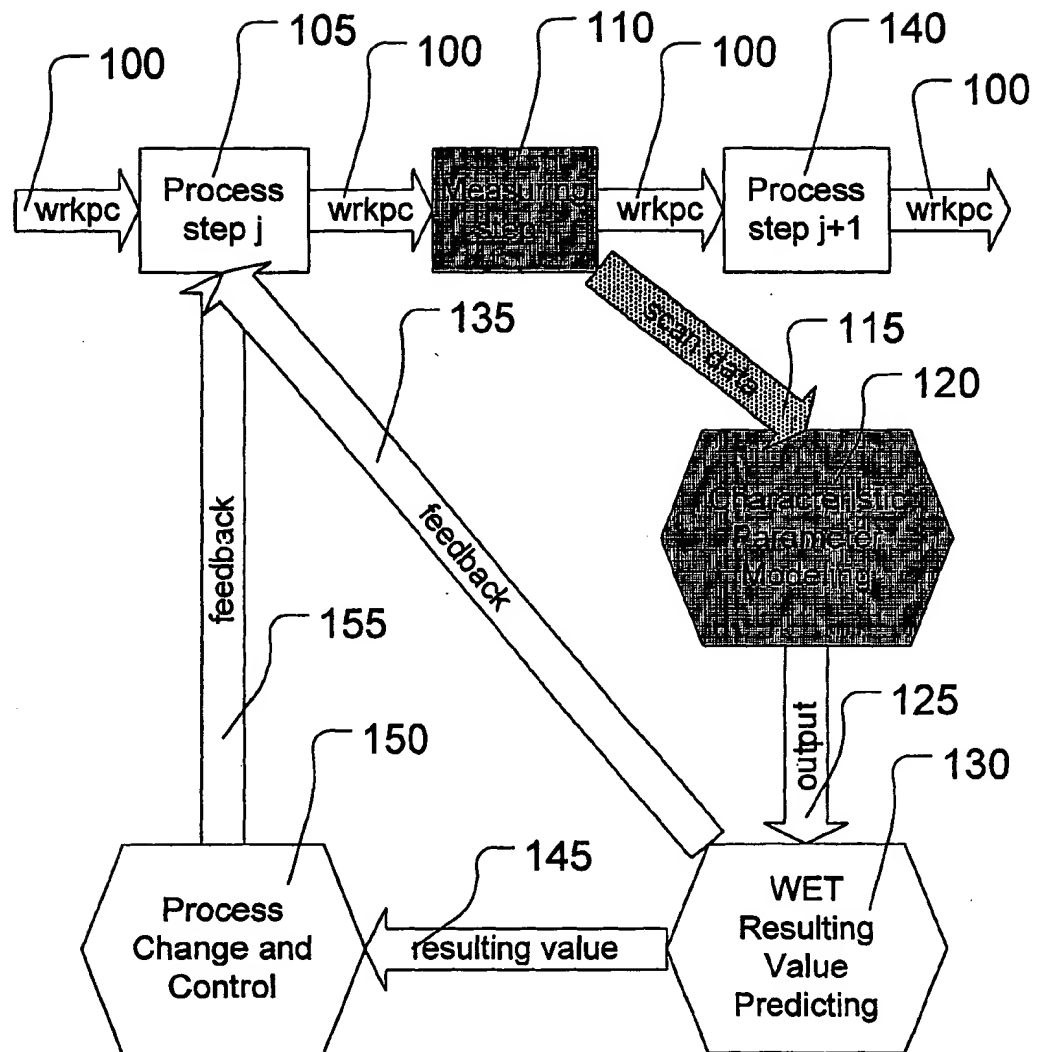
**Figure 1**

2 / 11

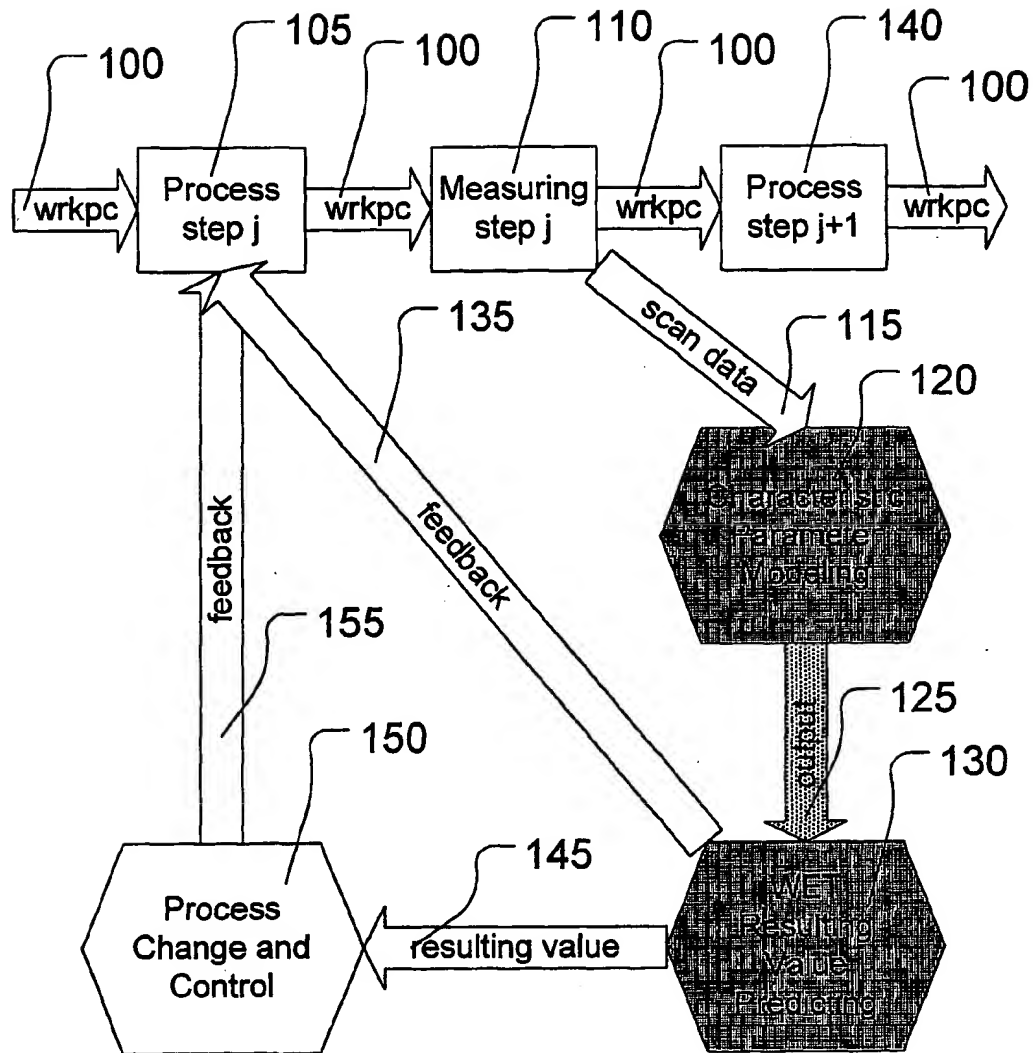
**Figure 2**

**Figure 3****Figure 4**

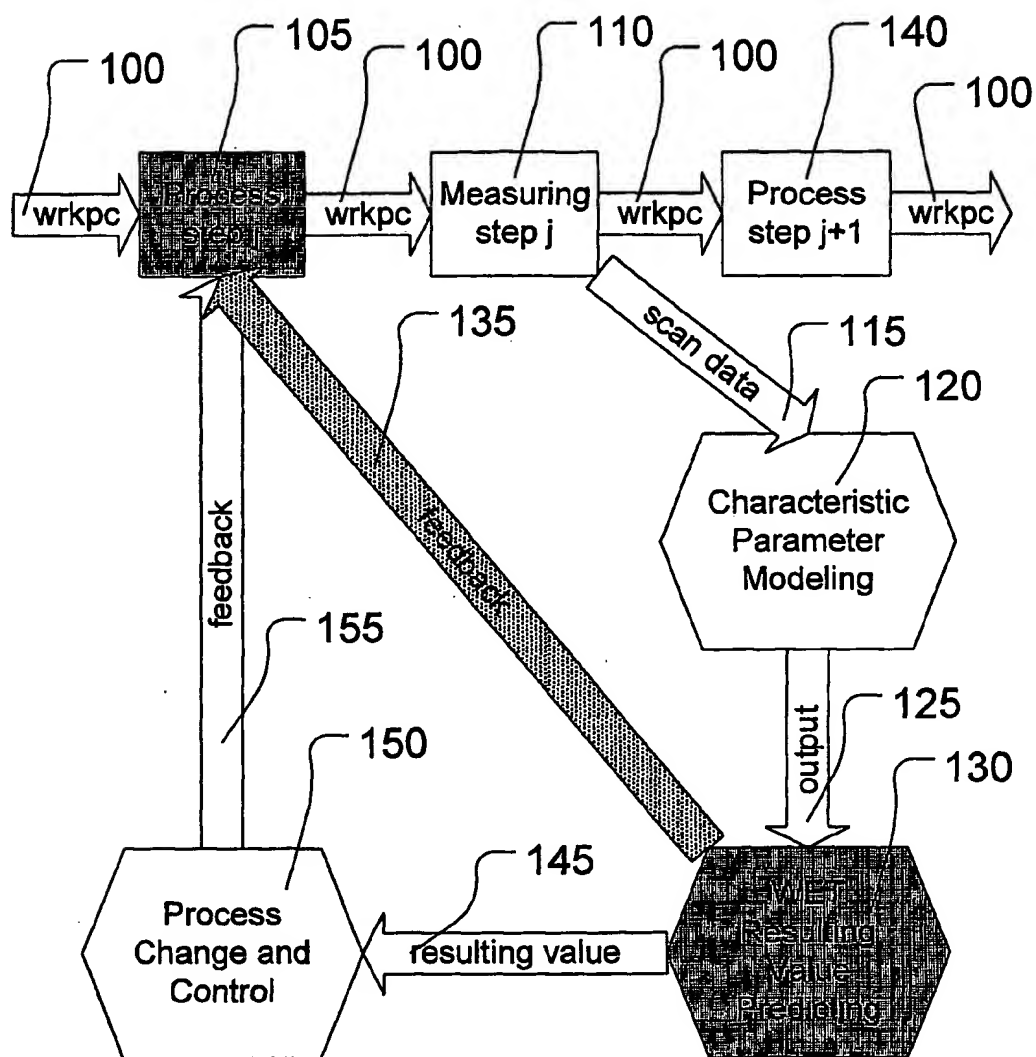
4 / 11

**Figure 5**

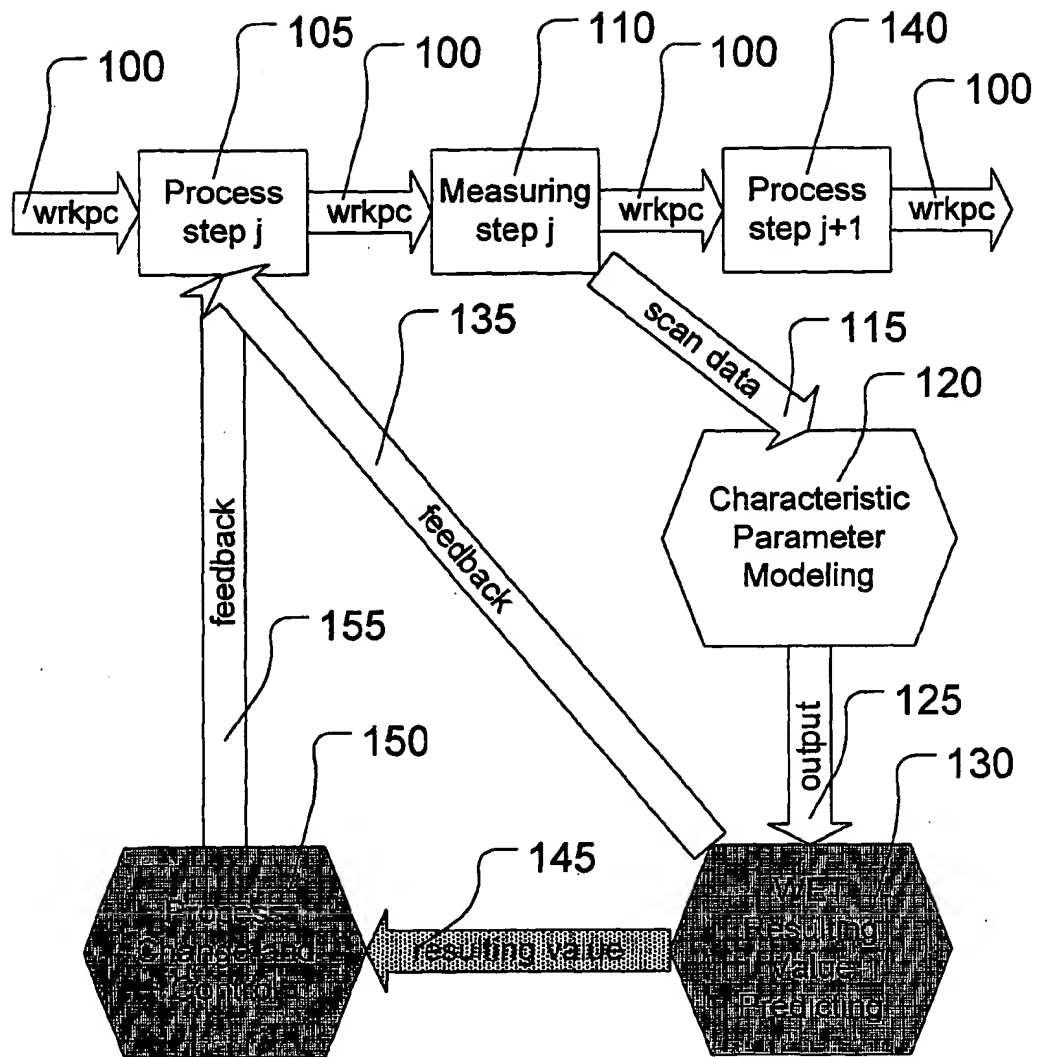
5 / 11

**Figure 6**

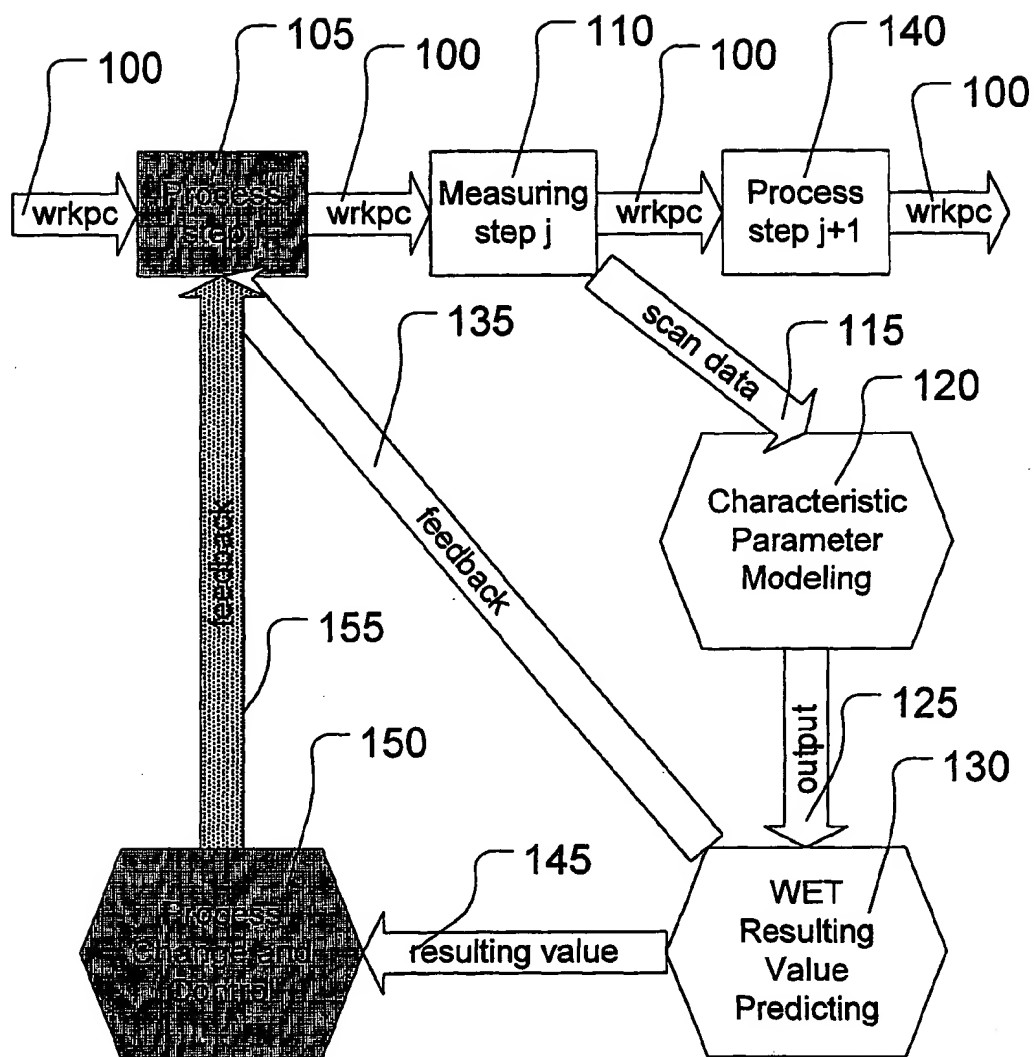
6 / 11

**Figure 7**

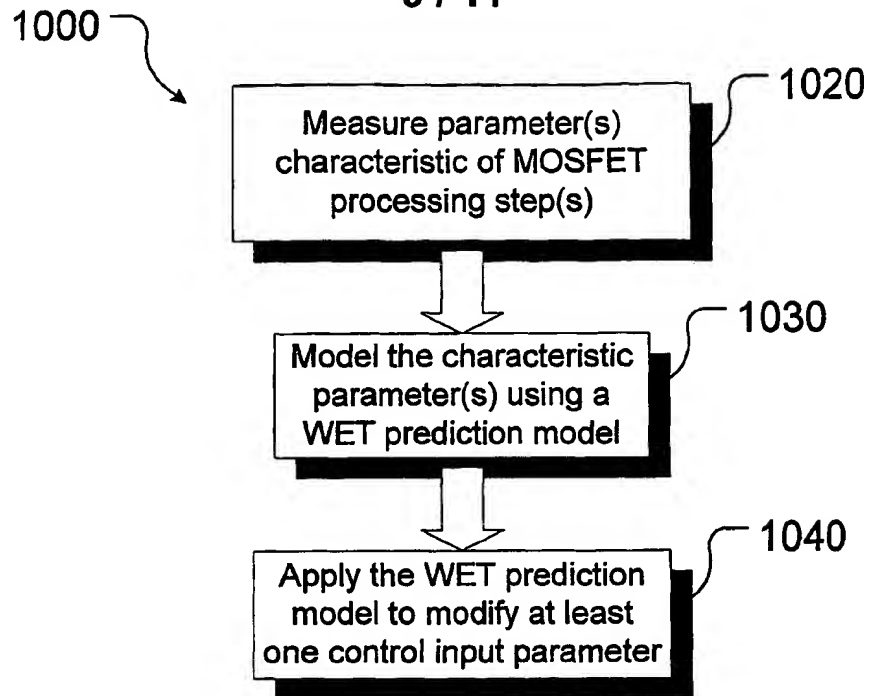
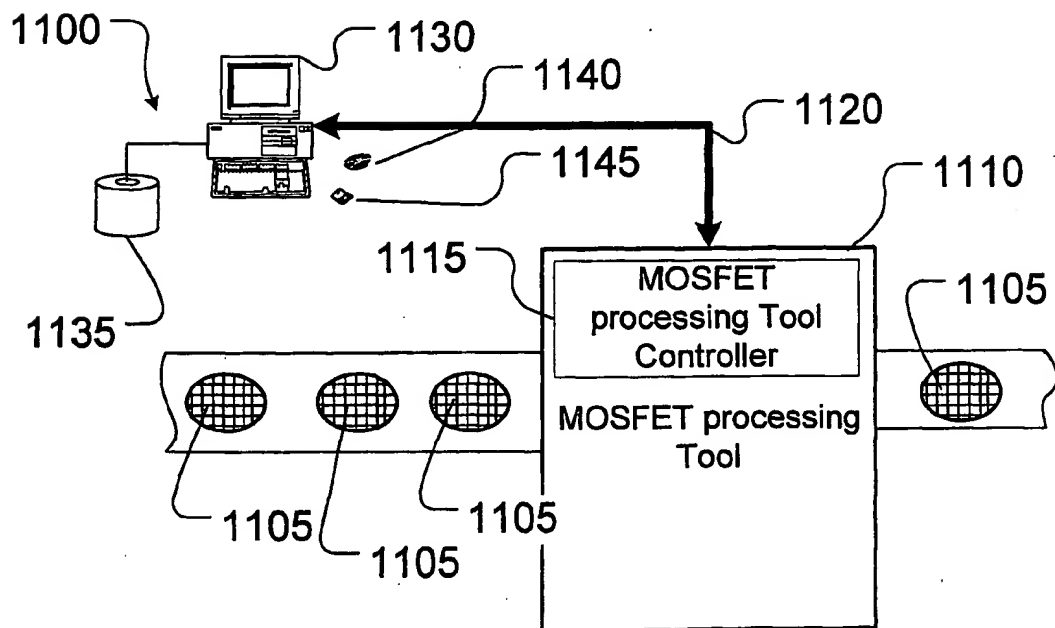
7 / 11

**Figure 8**

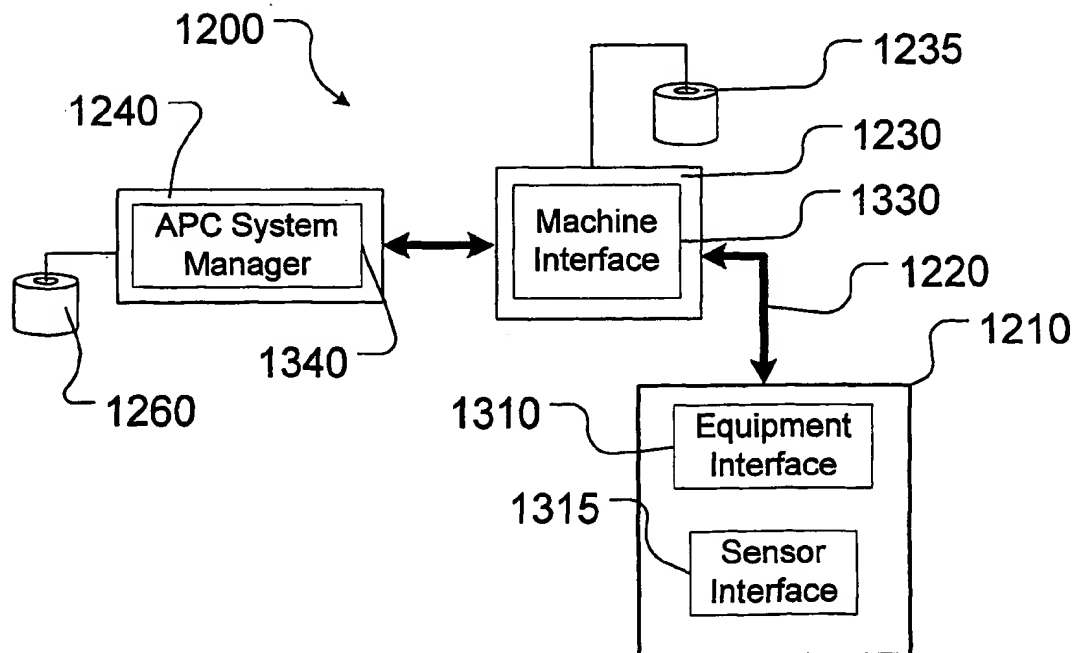
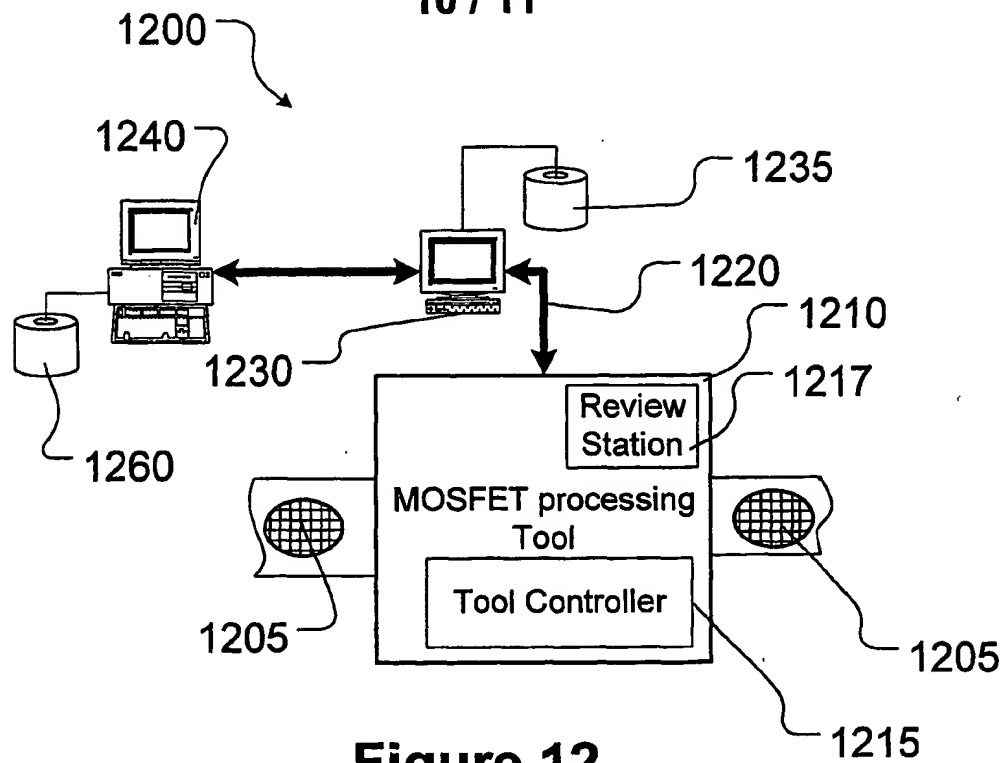
8 / 11

**Figure 9**

9 / 11

**Figure 10****Figure 11**

10 / 11



11 / 11

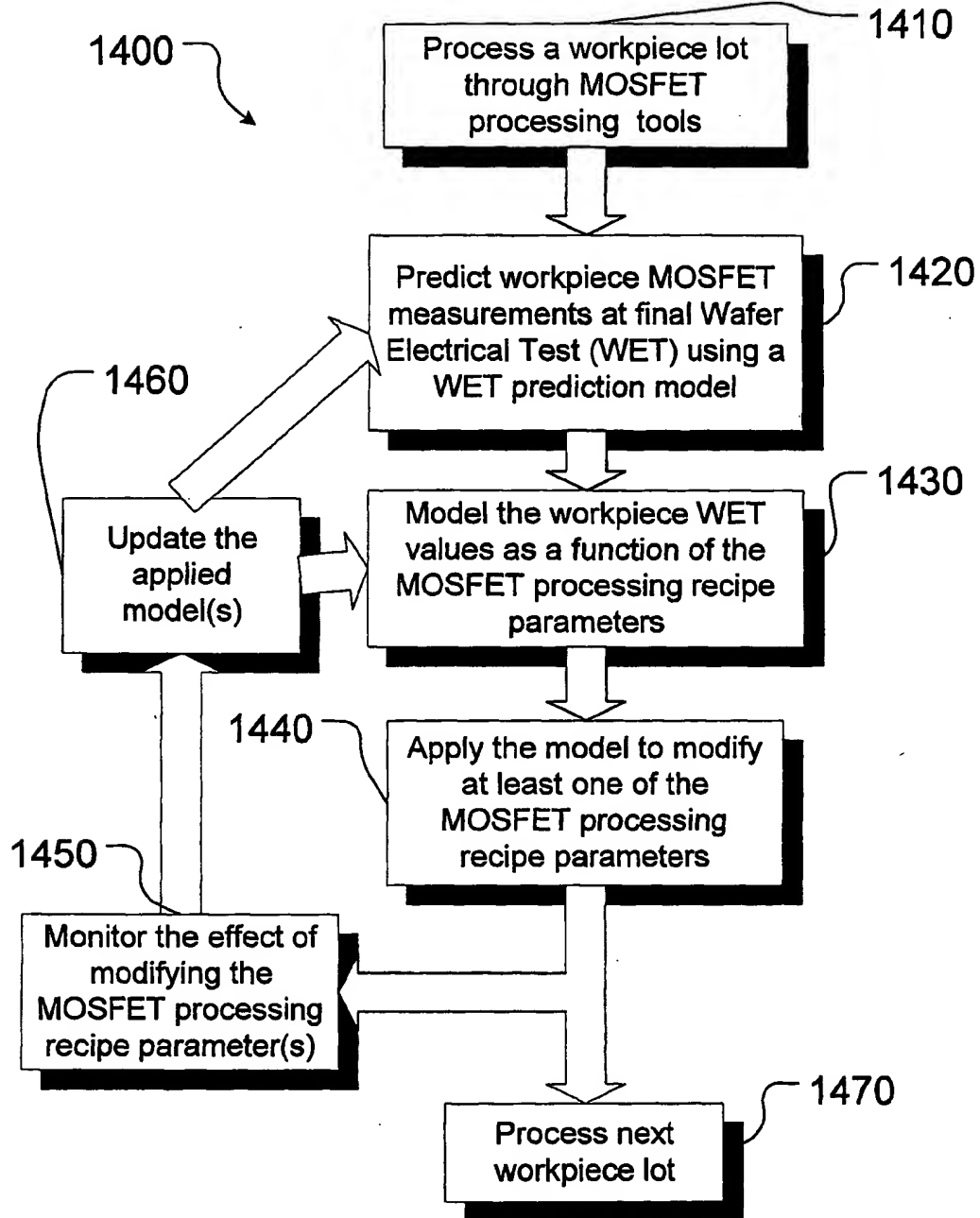


Figure 14